Kendall Atkinson
Weimin Han

TEXTS IN APPLIED MATHEMATICS 39

Theoretical Numerical Analysis

A Functional Analysis Framework

Third Edition

Springer

Texts in Applied Mathematics 39

Editors
J.E. Marsden
L. Sirovich
S.S. Antman

Advisors
G. Iooss
P. Holmes
D. Barkley
M. Dellnitz
P. Newton

For other volumes published in this series, go to
www.springer.com/series/1214

Kendall Atkinson · Weimin Han

Theoretical Numerical Analysis

A Functional Analysis Framework

Third Edition

Springer

Kendall Atkinson
Departments of Mathematics &
Computer Science
University of Iowa
Iowa City, IA 52242
USA
kendall-atkinson@uiowa.edu

Weimin Han
Department of Mathematics
University of Iowa
Iowa City, IA 52242
USA
whan@math.uiowa.edu

Series Editors

J.E. Marsden
Control and Dynamical Systems
107-81 California Institute of Technology
Pasadena, CA 91125
USA
marsden@cds.caltech.edu

L. Sirovich
Laboratory of Applied Mathematics
Department of Biomathematics
Mt. Sinai School of Medicine
Box 1012
New York, NY 10029-6574
USA
lawrence.sirovich@mssm.edu

S.S. Antman
Department of Mathematics
and
Institute for Physical Science
and Technology
University of Maryland
College Park, MD 20742-4015
USA
ssa@math.umd.edu

ISSN 0939-2475
ISBN 978-1-4419-0457-7 e-ISBN 978-1-4419-0458-4
DOI 10.1007/978-1-4419-0458-4
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009926473

Mathematics Subject Classification (2000): 65-01, 65-XX

© Springer Science+Business Media, LLC 2009
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to

Daisy and Clyde Atkinson
Hazel and Wray Fleming

and

Daqing Han, Suzhen Qin
Huidi Tang, Elizabeth and Michael

Series Preface

Mathematics is playing an ever more important role in the physical and biological sciences, provoking a blurring of boundaries between scientific disciplines and a resurgence of interest in the modern as well as the classical techniques of applied mathematics. This renewal of interest, both in research and teaching, has led to the establishment of the series: *Texts in Applied Mathematics* (*TAM*).

The development of new courses is a natural consequence of a high level of excitement on the research frontier as newer techniques, such as numerical and symbolic computer systems, dynamical systems, and chaos, mix with and reinforce the traditional methods of applied mathematics. Thus, the purpose of this textbook series is to meet the current and future needs of these advances and to encourage the teaching of new courses.

TAM will publish textbooks suitable for use in advanced undergraduate and beginning graduate courses, and will complement the *Applied Mathematical Sciences* (*AMS*) series, which will focus on advanced textbooks and research-level monographs.

Pasadena, California — J.E. Marsden
Providence, Rhode Island — L. Sirovich
College Park, Maryland — S.S. Antman

Preface

This textbook has grown out of a course which we teach periodically at the University of Iowa. We have beginning graduate students in mathematics who wish to work in numerical analysis from a theoretical perspective, and they need a background in those "tools of the trade" which we cover in this text. In the past, such students would ordinarily begin with a one-year course in *real and complex analysis*, followed by a one or two semester course in *functional analysis* and possibly a graduate level course in *ordinary differential equations, partial differential equations*, or *integral equations*. We still expect our students to take most of these standard courses. The course based on this book allows these students to move more rapidly into a research program.

The textbook covers basic results of functional analysis, approximation theory, Fourier analysis and wavelets, calculus and iteration methods for nonlinear equations, finite difference methods, Sobolev spaces and weak formulations of boundary value problems, finite element methods, elliptic variational inequalities and their numerical solution, numerical methods for solving integral equations of the second kind, boundary integral equations for planar regions with a smooth boundary curve, and multivariable polynomial approximations. The presentation of each topic is meant to be an introduction with a certain degree of depth. Comprehensive references on a particular topic are listed at the end of each chapter for further reading and study. For this third edition, we add a chapter on multivariable polynomial approximation and we revise numerous sections from the second edition to varying degrees. A good number of new exercises are included.

The material in the text is presented in a mixed manner. Some topics are treated with complete rigour, whereas others are simply presented without proof and perhaps illustrated (e.g. the principle of uniform boundedness). We have chosen to avoid introducing a formalized framework for *Lebesgue measure and integration* and also for *distribution theory*. Instead we use standard results on the completion of normed spaces and the unique extension of densely defined bounded linear operators. This permits us to introduce the Lebesgue spaces formally and without their concrete realization using measure theory. We describe some of the standard material on measure theory and distribution theory in an intuitive manner, believing this is sufficient for much of the subsequent mathematical development. In addition, we give a number of deeper results without proof, citing the existing literature. Examples of this are the *open mapping theorem, Hahn-Banach theorem, the principle of uniform boundedness*, and a number of the results on *Sobolev spaces*.

The choice of topics has been shaped by our research program and interests at the University of Iowa. These topics are important elsewhere, and we believe this text will be useful to students at other universities as well.

The book is divided into chapters, sections, and subsections as appropriate. Mathematical relations (equalities and inequalities) are numbered by chapter, section and their order of occurrence. For example, (1.2.3) is the third numbered mathematical relation in Section 1.2 of Chapter 1. Definitions, examples, theorems, lemmas, propositions, corollaries and remarks are numbered consecutively within each section, by chapter and section. For example, in Section 1.1, Definition 1.1.1 is followed by an example labeled as Example 1.1.2.

We give exercises at the end of most sections. The exercises are numbered consecutively by chapter and section. At the end of each chapter, we provide some short discussions of the literature, including recommendations for additional reading.

During the preparation of the book, we received helpful suggestions from numerous colleagues and friends. We particularly thank P.G. Ciarlet, William A. Kirk, Wenbin Liu, and David Stewart for the first edition, B. Bialecki, R. Glowinski, and A.J. Meir for the second edition, and Yuan Xu for the third edition. It is a pleasure to acknowledge the skillful editorial assistance from the Series Editor, Achi Dosanjh.

Contents

Series Preface **vii**

Preface **ix**

1 Linear Spaces **1**
1.1 Linear spaces 1
1.2 Normed spaces 7
1.2.1 Convergence 10
1.2.2 Banach spaces 13
1.2.3 Completion of normed spaces 15
1.3 Inner product spaces 22
1.3.1 Hilbert spaces 27
1.3.2 Orthogonality 28
1.4 Spaces of continuously differentiable functions 39
1.4.1 Hölder spaces 41
1.5 L^p spaces 44
1.6 Compact sets 49

2 Linear Operators on Normed Spaces **51**
2.1 Operators 52
2.2 Continuous linear operators 55
2.2.1 $\mathcal{L}(V, W)$ as a Banach space 59
2.3 The geometric series theorem and its variants 60
2.3.1 A generalization 64

2.3.2 A perturbation result 66
2.4 Some more results on linear operators 72
2.4.1 An extension theorem 72
2.4.2 Open mapping theorem 74
2.4.3 Principle of uniform boundedness 75
2.4.4 Convergence of numerical quadratures 76
2.5 Linear functionals 79
2.5.1 An extension theorem for linear functionals 80
2.5.2 The Riesz representation theorem 82
2.6 Adjoint operators 85
2.7 Weak convergence and weak compactness 90
2.8 Compact linear operators 95
2.8.1 Compact integral operators on $C(D)$ 96
2.8.2 Properties of compact operators 97
2.8.3 Integral operators on $L^2(a, b)$ 99
2.8.4 The Fredholm alternative theorem 101
2.8.5 Additional results on Fredholm integral equations . 105
2.9 The resolvent operator 109
2.9.1 $R(\lambda)$ as a holomorphic function 110

3 Approximation Theory 115
3.1 Approximation of continuous functions by polynomials . . . 116
3.2 Interpolation theory 118
3.2.1 Lagrange polynomial interpolation 120
3.2.2 Hermite polynomial interpolation 122
3.2.3 Piecewise polynomial interpolation 124
3.2.4 Trigonometric interpolation 126
3.3 Best approximation 131
3.3.1 Convexity, lower semicontinuity 132
3.3.2 Some abstract existence results 134
3.3.3 Existence of best approximation 137
3.3.4 Uniqueness of best approximation 138
3.4 Best approximations in inner product spaces, projection on closed convex sets 142
3.5 Orthogonal polynomials 149
3.6 Projection operators 154
3.7 Uniform error bounds 157
3.7.1 Uniform error bounds for L^2-approximations 160
3.7.2 L^2-approximations using polynomials 162
3.7.3 Interpolatory projections and their convergence . . . 164

4 Fourier Analysis and Wavelets 167
4.1 Fourier series 167
4.2 Fourier transform 181
4.3 Discrete Fourier transform 187

4.4 Haar wavelets 191
4.5 Multiresolution analysis 199

5 Nonlinear Equations and Their Solution by Iteration 207
5.1 The Banach fixed-point theorem 208
5.2 Applications to iterative methods 212
5.2.1 Nonlinear algebraic equations 213
5.2.2 Linear algebraic systems 214
5.2.3 Linear and nonlinear integral equations 216
5.2.4 Ordinary differential equations in Banach spaces . . 221
5.3 Differential calculus for nonlinear operators 225
5.3.1 Fréchet and Gâteaux derivatives 225
5.3.2 Mean value theorems 229
5.3.3 Partial derivatives 230
5.3.4 The Gâteaux derivative and convex minimization . . 231
5.4 Newton's method 236
5.4.1 Newton's method in Banach spaces 236
5.4.2 Applications 239
5.5 Completely continuous vector fields 241
5.5.1 The rotation of a completely continuous vector field 243
5.6 Conjugate gradient method for operator equations 245

6 Finite Difference Method 253
6.1 Finite difference approximations 253
6.2 Lax equivalence theorem 260
6.3 More on convergence 269

7 Sobolev Spaces 277
7.1 Weak derivatives 277
7.2 Sobolev spaces 283
7.2.1 Sobolev spaces of integer order 284
7.2.2 Sobolev spaces of real order 290
7.2.3 Sobolev spaces over boundaries 292
7.3 Properties 293
7.3.1 Approximation by smooth functions 293
7.3.2 Extensions 294
7.3.3 Sobolev embedding theorems 295
7.3.4 Traces 297
7.3.5 Equivalent norms 298
7.3.6 A Sobolev quotient space 302
7.4 Characterization of Sobolev spaces via the Fourier transform 308
7.5 Periodic Sobolev spaces 311
7.5.1 The dual space 314
7.5.2 Embedding results 315
7.5.3 Approximation results 316

7.5.4 An illustrative example of an operator 317
7.5.5 Spherical polynomials and spherical harmonics . . . 318
7.6 Integration by parts formulas 323

8 Weak Formulations of Elliptic Boundary Value Problems 327
8.1 A model boundary value problem 328
8.2 Some general results on existence and uniqueness 330
8.3 The Lax-Milgram Lemma 334
8.4 Weak formulations of linear elliptic boundary value problems 338
8.4.1 Problems with homogeneous Dirichlet boundary conditions 338
8.4.2 Problems with non-homogeneous Dirichlet boundary conditions 339
8.4.3 Problems with Neumann boundary conditions 341
8.4.4 Problems with mixed boundary conditions 343
8.4.5 A general linear second-order elliptic boundary value problem 344
8.5 A boundary value problem of linearized elasticity 348
8.6 Mixed and dual formulations 354
8.7 Generalized Lax-Milgram Lemma 359
8.8 A nonlinear problem 361

9 The Galerkin Method and Its Variants 367
9.1 The Galerkin method 367
9.2 The Petrov-Galerkin method 374
9.3 Generalized Galerkin method 376
9.4 Conjugate gradient method: variational formulation 378

10 Finite Element Analysis 383
10.1 One-dimensional examples 384
10.1.1 Linear elements for a second-order problem 384
10.1.2 High order elements and the condensation technique 389
10.1.3 Reference element technique 390
10.2 Basics of the finite element method 393
10.2.1 Continuous linear elements 394
10.2.2 Affine-equivalent finite elements 400
10.2.3 Finite element spaces 404
10.3 Error estimates of finite element interpolations 406
10.3.1 Local interpolations 407
10.3.2 Interpolation error estimates on the reference element 408
10.3.3 Local interpolation error estimates 409
10.3.4 Global interpolation error estimates 412
10.4 Convergence and error estimates 415

11 Elliptic Variational Inequalities and Their Numerical Approximations 423
11.1 From variational equations to variational inequalities 423
11.2 Existence and uniqueness based on convex minimization . . 428
11.3 Existence and uniqueness results for a family of EVIs 430
11.4 Numerical approximations 442
11.5 Some contact problems in elasticity 458
11.5.1 A frictional contact problem 460
11.5.2 A Signorini frictionless contact problem 465

12 Numerical Solution of Fredholm Integral Equations of the Second Kind 473
12.1 Projection methods: General theory 474
12.1.1 Collocation methods 474
12.1.2 Galerkin methods 476
12.1.3 A general theoretical framework 477
12.2 Examples 483
12.2.1 Piecewise linear collocation 483
12.2.2 Trigonometric polynomial collocation 486
12.2.3 A piecewise linear Galerkin method 488
12.2.4 A Galerkin method with trigonometric polynomials . 490
12.3 Iterated projection methods 494
12.3.1 The iterated Galerkin method 497
12.3.2 The iterated collocation solution 498
12.4 The Nyström method 504
12.4.1 The Nyström method for continuous kernel functions 505
12.4.2 Properties and error analysis of the Nyström method 507
12.4.3 Collectively compact operator approximations 516
12.5 Product integration 518
12.5.1 Error analysis 520
12.5.2 Generalizations to other kernel functions 523
12.5.3 Improved error results for special kernels 525
12.5.4 Product integration with graded meshes 525
12.5.5 The relationship of product integration and collocation methods 529
12.6 Iteration methods 531
12.6.1 A two-grid iteration method for the Nyström method 532
12.6.2 Convergence analysis 535
12.6.3 The iteration method for the linear system 538
12.6.4 An operations count 540
12.7 Projection methods for nonlinear equations 542
12.7.1 Linearization 542
12.7.2 A homotopy argument 545
12.7.3 The approximating finite-dimensional problem . . . 547

13 Boundary Integral Equations **551**
13.1 Boundary integral equations 552
13.1.1 Green's identities and representation formula 553
13.1.2 The Kelvin transformation and exterior problems . 555
13.1.3 Boundary integral equations of direct type 559
13.2 Boundary integral equations of the second kind 565
13.2.1 Evaluation of the double layer potential 568
13.2.2 The exterior Neumann problem 571
13.3 A boundary integral equation of the first kind 577
13.3.1 A numerical method 579

14 Multivariable Polynomial Approximations **583**
14.1 Notation and best approximation results 583
14.2 Orthogonal polynomials 585
14.2.1 Triple recursion relation 588
14.2.2 The orthogonal projection operator and its error . . 590
14.3 Hyperinterpolation 592
14.3.1 The norm of the hyperinterpolation operator 593
14.4 A Galerkin method for elliptic equations 593
14.4.1 The Galerkin method and its convergence 595

References **601**

Index **617**

1
Linear Spaces

Linear (or vector) spaces are the standard setting for studying and solving a large proportion of the problems in differential and integral equations, approximation theory, optimization theory, and other topics in applied mathematics. In this chapter, we gather together some concepts and results concerning various aspects of linear spaces, especially some of the more important linear spaces such as Banach spaces, Hilbert spaces, and certain function spaces that are used frequently in this work and in applied mathematics generally.

1.1 Linear spaces

A linear space is a set of elements equipped with two binary operations, called vector addition and scalar multiplication, in such a way that the operations behave linearly.

Definition 1.1.1 *Let V be a set of objects, to be called* vectors*; and let $\mathbb{K}$ be a set of scalars, either $\mathbb{R}$, the set of real numbers, or $\mathbb{C}$, the set of complex numbers. Assume there are two operations: $(u, v) \mapsto u + v \in V$ and $(\alpha, v) \mapsto \alpha v \in V$, called* addition *and* scalar multiplication *respectively, defined for any $u, v \in V$ and any $\alpha \in \mathbb{K}$. These operations are to satisfy the following rules.*

1. *$u + v = v + u$ for any $u, v \in V$ (commutative law);*
2. *$(u + v) + w = u + (v + w)$ for any $u, v, w \in V$ (associative law);*

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_1,
© Springer Science + Business Media, LLC 2009

3. *there is an element* $0 \in V$ *such that* $0+v = v$ *for any* $v \in V$ *(existence of the zero element);*

4. *for any* $v \in V$*, there is an element* $-v \in V$ *such that* $v + (-v) = 0$ *(existence of negative elements);*

5. $1v = v$ *for any* $v \in V$*;*

6. $\alpha(\beta v) = (\alpha\beta)v$ *for any* $v \in V$*, any* $\alpha, \beta \in \mathbb{K}$ *(associative law for scalar multiplication);*

7. $\alpha(u + v) = \alpha u + \alpha v$ *and* $(\alpha + \beta)v = \alpha v + \beta v$ *for any* $u, v \in V$*, and any* $\alpha, \beta \in \mathbb{K}$ *(distributive laws).*

Then V *is called a* linear space*, or a* vector space*.*

When $\mathbb{K}$ is the set of the real numbers, V is a real linear space; and when $\mathbb{K}$ is the set of the complex numbers, V becomes a complex linear space. In this work, most of the time we only deal with real linear spaces. So when we say V is a linear space, the reader should usually assume V is a real linear space, unless explicitly stated otherwise.

Some remarks are in order concerning the definition of a linear space. From the commutative law and the associative law, we observe that to add several elements, the order of summation does not matter, and it does not cause any ambiguity to write expressions such as $u + v + w$ or $\sum_{i=1}^{n} u_i$. By using the commutative law and the associative law, it is not difficult to verify that the zero element and the negative element $(-v)$ of a given element $v \in V$ are unique, and they can be equivalently defined through the relations $v + 0 = v$ for any $v \in V$, and $(-v) + v = 0$. Below, we write $u - v$ for $u + (-v)$. This defines the subtraction of two vectors. Sometimes, we will also refer to a vector as a point.

Example 1.1.2 (a) The set $\mathbb{R}$ of the real numbers is a real linear space when the addition and scalar multiplication are the usual addition and multiplication. Similarly, the set $\mathbb{C}$ of the complex numbers is a complex linear space.

(b) Let d be a positive integer. The letter d is used generally in this work for the spatial dimension. The set of all vectors with d real components, with the usual vector addition and scalar multiplication, forms a linear space $\mathbb{R}^d$. A typical element in $\mathbb{R}^d$ can be expressed as $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, where $x_1, \ldots, x_d \in \mathbb{R}$. Similarly, $\mathbb{C}^d$ is a complex linear space.

(c) Let $\Omega \subset \mathbb{R}^d$ be an open set of $\mathbb{R}^d$. In this work, the symbol Ω always stands for an open subset of $\mathbb{R}^d$. The set of all the continuous functions on Ω forms a linear space $C(\Omega)$, under the usual addition and scalar multiplication of functions: For $f, g \in C(\Omega)$, the function $f + g$ defined by

$$(f + g)(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega,$$

belongs to $C(\Omega)$, as does the scalar multiplication function αf defined through

$$(\alpha f)(\boldsymbol{x}) = \alpha\, f(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega.$$

Similarly, $C(\overline{\Omega})$ denotes the space of continuous functions on the closed set $\overline{\Omega}$. Clearly, any $C(\overline{\Omega})$ function is continuous on Ω, and thus can be viewed as a $C(\Omega)$ function. Conversely, if $f \in C(\Omega)$ is uniformly continuous on Ω and Ω is bounded, then f can be continuously extended to $\partial\Omega$, the boundary of Ω, and the extended function belongs to $C(\overline{\Omega})$. Recall that f defined on Ω is uniformly continuous if for any $\varepsilon > 0$, there exists a $\delta = \delta(f, \varepsilon) > 0$ such that

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| < \varepsilon$$

whenever $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ with $\|\boldsymbol{x} - \boldsymbol{y}\| < \delta$. Note that a $C(\Omega)$ function can behave badly near $\partial\Omega$; consider for example $f(x) = \sin(1/x)$, $0 < x < 1$, for x near 0.

(d) A related function space is $C(D)$, containing all functions $f : D \to \mathbb{K}$ which are continuous on a general set $D \subset \mathbb{R}^d$. The arbitrary set D can be an open or closed set in $\mathbb{R}^d$, or perhaps neither; and it can be a lower dimensional set such as a portion of the boundary of an open set in $\mathbb{R}^d$. When D is a closed and bounded subset of $\mathbb{R}^d$, a function from the space $C(D)$ is necessarily bounded.

(e) For any non-negative integer m, we may define the space $C^m(\Omega)$ as the set of all the functions, which together with their derivatives of orders up to m are continuous on Ω. We may also define $C^m(\overline{\Omega})$ to be the space of all the functions, which together with their derivatives of orders up to m are continuous on $\overline{\Omega}$. These function spaces are discussed at length in Section 1.4.

(f) The space of continuous 2π-periodic functions is denoted by $C_p(2\pi)$. It is the set of all $f \in C(-\infty, \infty)$ for which

$$f(x + 2\pi) = f(x), \quad -\infty < x < \infty.$$

For an integer $k \geq 0$, the space $C_p^k(2\pi)$ denotes the set of all functions in $C_p(2\pi)$ which have k continuous derivatives on $(-\infty, \infty)$. We usually write $C_p^0(2\pi)$ as simply $C_p(2\pi)$. These spaces are used in connection with problems in which periodicity plays a major role. □

Definition 1.1.3 *A* subspace W *of the linear space* V *is a subset of* V *which is closed under the addition and scalar multiplication operations of* V*, i.e., for any* $u, v \in W$ *and any* $\alpha \in \mathbb{K}$*, we have* $u + v \in W$ *and* $\alpha v \in W$*.*

It can be verified that W itself, equipped with the addition and scalar multiplication operations of V, is a linear space.

Example 1.1.4 In the linear space $\mathbb{R}^3$,

$$W = \{\boldsymbol{x} = (x_1, x_2, 0)^T \mid x_1, x_2 \in \mathbb{R}\}$$

is a subspace, consisting of all the vectors on the x_1x_2-plane. In contrast,

$$\widehat{W} = \{\boldsymbol{x} = (x_1, x_2, 1)^T \mid x_1, x_2 \in \mathbb{R}\}$$

is not a subspace. Nevertheless, we observe that $\widehat{W}$ is a translation of the subspace W,

$$\widehat{W} = \boldsymbol{x}_0 + W$$

where $\boldsymbol{x}_0 = (0, 0, 1)^T$. The set $\widehat{W}$ is an example of an *affine* set. □

Given vectors $v_1, \ldots, v_n \in V$ and scalars $\alpha_1, \ldots, \alpha_n \in \mathbb{K}$, we call

$$\sum_{i=1}^{n} \alpha_i v_i = \alpha_1 v_1 + \cdots + \alpha_n v_n$$

a *linear combination* of $v_1, \ldots, v_n$. It is meaningful to remove "redundant" vectors from the linear combination. Thus we introduce the concepts of linear dependence and independence.

Definition 1.1.5 *We say $v_1, \ldots, v_n \in V$ are* linearly dependent *if there are scalars $\alpha_i \in \mathbb{K}$, $1 \le i \le n$, with at least one α_i nonzero such that*

$$\sum_{i=1}^{n} \alpha_i v_i = 0. \tag{1.1.1}$$

We say $v_1, \ldots, v_n \in V$ are linearly independent *if they are not linearly dependent, in other words, if* (1.1.1) *implies $\alpha_i = 0$ for $i = 1, 2, \ldots, n$.*

We observe that $v_1, \ldots, v_n$ are linearly dependent if and only if at least one of the vectors can be expressed as a linear combination of the rest of the vectors. In particular, a set of vectors containing the zero element is always linearly dependent. Similarly, $v_1, \ldots, v_n$ are linearly independent if and only if none of the vectors can be expressed as a linear combination of the rest of the vectors; in other words, none of the vectors is "redundant".

Example 1.1.6 In $\mathbb{R}^d$, d vectors $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)})^T$, $1 \le i \le d$, are linearly independent if and only if the determinant

$$\begin{vmatrix} x_1^{(1)} & \cdots & x_1^{(d)} \\ \vdots & \ddots & \vdots \\ x_d^{(1)} & \cdots & x_d^{(d)} \end{vmatrix}$$

is nonzero. This follows from a standard result in linear algebra. The condition (1.1.1) is equivalent to a homogeneous system of linear equations, and a standard result of linear algebra says that this system has $(0, \ldots, 0)^T$ as its only solution if and only if the above determinant is nonzero. □

Example 1.1.7 Within the space $C[0,1]$, the vectors $1, x, x^2, \ldots, x^n$ are linearly independent. This can be proved in several ways. Assuming

$$\sum_{j=0}^{n} \alpha_j x^j = 0, \quad 0 \le x \le 1,$$

we can form its first n derivatives. Setting $x = 0$ in this polynomial and its derivatives will lead to $\alpha_j = 0$ for $j = 0, 1, \ldots, n$. □

Definition 1.1.8 *The* span *of $v_1, \ldots, v_n \in V$ is defined to be the set of all the linear combinations of these vectors:*

$$\operatorname{span}\{v_1, \ldots, v_n\} = \left\{ \sum_{i=1}^{n} \alpha_i v_i \;\middle|\; \alpha_i \in \mathbb{K},\ 1 \le i \le n \right\}.$$

Evidently, $\operatorname{span}\{v_1, \ldots, v_n\}$ is a linear subspace of V. Most of the time, we apply this definition for the case where $v_1, \ldots, v_n$ are linearly independent.

Definition 1.1.9 *A linear space V is said to be* finite dimensional *if there exists a finite maximal set of independent vectors $\{v_1, \ldots, v_n\}$; i.e., the set $\{v_1, \ldots, v_n\}$ is linearly independent, but $\{v_1, \ldots, v_n, v_{n+1}\}$ is linearly dependent for any $v_{n+1} \in V$. The set $\{v_1, \ldots, v_n\}$ is called a* basis *of the space. If such a finite basis for V does not exist, then V is said to be* infinite dimensional.

We see that a basis is a set of independent vectors such that any vector in the space can be written as a linear combination of them. Obviously a basis is not unique, yet we have the following important result.

Theorem 1.1.10 *For a finite dimensional linear space, every basis for V contains exactly the same number of vectors. This number is called the* dimension *of the space, denoted by* $\dim V$.

A proof of this result can be found in most introductory textbooks on linear algebra; for example, see [6, Section 5.4].

Example 1.1.11 The space $\mathbb{R}^d$ is d-dimensional. There are infinitely many possible choices for a basis of the space. A canonical basis for this space is $\{\boldsymbol{e}_i\}_{i=1}^{d}$, where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ in which the single 1 is in component i. □

Example 1.1.12 In the space $\mathbb{P}_n$ of the polynomials of degree less than or equal to n, $\{1, x, \ldots, x^n\}$ is a basis and we have $\dim(\mathbb{P}_n) = n + 1$. In the subspace

$$\mathbb{P}_{n,0} = \{p \in \mathbb{P}_n \mid p(0) = p(1) = 0\},$$

a basis is given by the functions $x\,(1-x)$, $x^2(1-x)$, $\ldots$, $x^{n-1}(1-x)$. We observe that

$$\dim(\mathbb{P}_{n,0}) = \dim(\mathbb{P}_n) - 2.$$

The difference 2 in the dimensions reflects the two zero value conditions at 0 and 1 in the definition of $\mathbb{P}_{n,0}$. □

We now introduce the concept of a linear function.

Definition 1.1.13 *Let L be a function from one linear space V to another linear space W. We say L is a linear function if*
(a) *for all $u, v \in V$,*

$$L(u+v) = L(u) + L(v);$$

(b) *for all $v \in V$ and all $\alpha \in \mathbb{K}$,*

$$L(\alpha v) = \alpha L(v).$$

For such a linear function, we often write $L(v)$ for Lv.

This definition is extended and discussed extensively in Chapter 2. Other common names are *linear mapping*, *linear operator*, and *linear transformation.*

Definition 1.1.14 *Two linear spaces U and V are said to be* isomorphic*, if there is a linear bijective (i.e., one-to-one and onto) function $\ell : U \to V$.*

Many properties of a linear space U hold for any other linear space V that is isomorphic to U; and then the explicit contents of the space do not matter in the analysis of these properties. This usually proves to be convenient. One such example is that if U and V are isomorphic and are finite dimensional, then their dimensions are equal, a basis of V can be obtained from that of U by applying the mapping ℓ, and a basis of U can be obtained from that of V by applying the inverse mapping of ℓ.

Example 1.1.15 The set $\mathbb{P}_k$ of all polynomials of degree less than or equal to k is a subspace of continuous function space $C[0,1]$. An element in the space $\mathbb{P}_k$ has the form $a_0 + a_1x + \cdots + a_kx^k$. The mapping $\ell : a_0 + a_1x + \cdots + a_kx^k \mapsto (a_0, a_1, \ldots, a_k)^T$ is bijective from $\mathbb{P}_k$ to $\mathbb{R}^{k+1}$. Thus, $\mathbb{P}_k$ is isomorphic to $\mathbb{R}^{k+1}$. □

Definition 1.1.16 *Let U and V be two linear spaces. The Cartesian product of the spaces, $W = U \times V$, is defined by*

$$W = \{w = (u, v) \mid u \in U,\ v \in V\}$$

endowed with componentwise addition and scalar multiplication

$$(u_1, v_1) + (u_2, v_2) = (u_1 + u_2, v_1 + v_2) \quad \forall\, (u_1, v_1), (u_2, v_2) \in W,$$
$$\alpha\,(u, v) = (\alpha\, u, \alpha\, v) \quad \forall\, (u, v) \in W,\ \forall\, \alpha \in \mathbb{K}\,.$$

It is easy to verify that W is a linear space. The definition can be extended in a straightforward way for the Cartesian product of any finite number of linear spaces.

Example 1.1.17 The real plane can be viewed as the *Cartesian product* of two real lines: $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. In general,

$$\mathbb{R}^d = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{d \text{ times}}. \qquad \square$$

Exercise 1.1.1 Show that the set of all continuous solutions of the differential equation $u''(x) + u(x) = 0$ is a finite-dimensional linear space. Is the set of all continuous solutions of $u''(x) + u(x) = 1$ a linear space?

Exercise 1.1.2 When is the set $\{v \in C[0,1] \mid v(0) = a\}$ a linear space?

Exercise 1.1.3 Show that in any linear space V, a set of vectors is always linearly dependent if one of the vectors is zero.

Exercise 1.1.4 Let $\{v_1, \ldots, v_n\}$ be a basis of an n-dimensional space V. Show that for any $v \in V$, there are scalars $\alpha_1, \ldots, \alpha_n$ such that

$$v = \sum_{i=1}^{n} \alpha_i v_i,$$

and the scalars $\alpha_1, \ldots, \alpha_n$ are uniquely determined by v.

Exercise 1.1.5 Assume U and V are finite dimensional linear spaces, and let $\{u_1, \ldots, u_n\}$ and $\{v_1, \ldots, v_m\}$ be bases for them, respectively. Using these bases, create a basis for $W = U \times V$. Determine $\dim W$.

1.2 Normed spaces

The previous section is devoted to the algebraic structure of spaces. In this section, we turn to the topological structure of spaces. In numerical analysis, we need to frequently examine the closeness of a numerical solution to the exact solution. To answer the question quantitatively, we need to have a measure on the magnitude of the difference between the numerical solution and the exact solution. A norm of a vector in a linear space provides such a measure.

Definition 1.2.1 *Given a linear space V, a* norm $\|\cdot\|$ *is a function from V to $\mathbb{R}$ with the following properties.*

1. $\|v\| \geq 0$ *for any* $v \in V$, *and* $\|v\| = 0$ *if and only if* $v = 0$;

2. $\|\alpha v\| = |\alpha|\,\|v\|$ *for any* $v \in V$ *and* $\alpha \in \mathbb{K}$;

3. $\|u + v\| \le \|u\| + \|v\|$ *for any* $u, v \in V$.

The space V *equipped with the norm* $\|\cdot\|$, $(V, \|\cdot\|)$, *is called a* normed linear space *or a* normed space. *We usually say* V *is a normed space when the definition of the norm is clear from the context.*

Some remarks are in order on the definition of a norm. The three axioms in the definition mimic the principal properties of the notion of the ordinary length of a vector in $\mathbb{R}^2$ or $\mathbb{R}^3$. The first axiom says the norm of any vector must be non-negative, and the only vector with zero norm is zero. The second axiom is usually called *positive homogeneity*. The third axiom is also called the *triangle inequality*, which is a direct extension of the triangle inequality on the plane: The length of any side of a triangle is bounded by the sum of the lengths of the other two sides. With the definition of a norm, we can use the quantity $\|u - v\|$ as a measure for the distance between u and v.

Definition 1.2.2 *Given a linear space* V, *a* semi-norm $|\cdot|$ *is a function from* V *to* $\mathbb{R}$ *with the properties of a norm except that* $|v| = 0$ *does not necessarily imply* $v = 0$.

One place in this work where the notion of a semi-norm plays an important role is in estimating the error of polynomial interpolation.

Example 1.2.3 (a) For $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, the formula

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{d} x_i^2 \right)^{1/2} \tag{1.2.1}$$

defines a norm in the space $\mathbb{R}^d$ (Exercise 1.2.6), called the *Euclidean norm*, which is the usual norm for the space $\mathbb{R}^d$. When $d = 1$, the norm coincides with the absolute value: $\|x\|_2 = |x|$ for $x \in \mathbb{R}$.
(b) More generally, for $1 \le p \le \infty$, the formulas

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} \qquad \text{for } 1 \le p < \infty, \tag{1.2.2}$$

$$\|\boldsymbol{x}\|_\infty = \max_{1 \le i \le d} |x_i| \tag{1.2.3}$$

define norms in the space $\mathbb{R}^d$ (see Exercise 1.2.6 for $p = 1, 2, \infty$, and Exercise 1.5.7 for other values of p). The norm $\|\cdot\|_p$ is called the *p-norm*, and $\|\cdot\|_\infty$ is called the *maximum* or *infinity norm*. It can be shown that

$$\|\boldsymbol{x}\|_\infty = \lim_{p \to \infty} \|\boldsymbol{x}\|_p$$

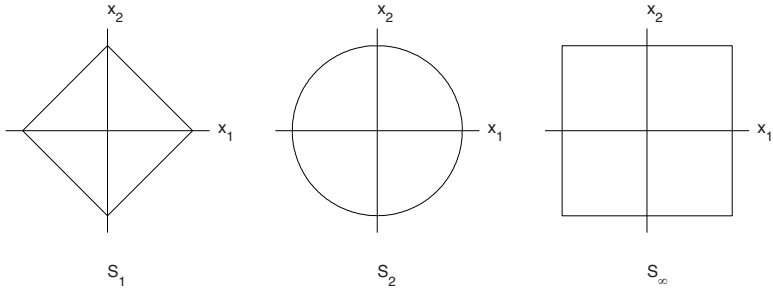

FIGURE 1.1. The unit circle $S_p = \{\boldsymbol{x} \in \mathbb{R}^2 \mid \|\boldsymbol{x}\|_p = 1\}$ for $p = 1, 2, \infty$

either directly or by using the inequality (1.2.6) given below. Again, when $d = 1$, all these norms coincide with the absolute value: $\|x\|_p = |x|$, $x \in \mathbb{R}$. Over $\mathbb{R}^d$, the most commonly used norms are $\|\cdot\|_p$, $p = 1, 2, \infty$. The unit circle in $\mathbb{R}^2$ for each of these norms is shown in Figure 1.2. □

Example 1.2.4 For $p \in [1, \infty]$, the space ℓ^p is defined as

$$\ell^p = \{v = (v_n)_{n\geq 1} \mid \|v\|_{\ell^p} < \infty\} \tag{1.2.4}$$

with the norm

$$\|v\|_{\ell^p} = \begin{cases} \left(\sum_{n=1}^{\infty} |v_n|^p\right)^{1/p} & \text{if } p < \infty, \\ \sup_{n\geq 1} |v_n| & \text{if } p = \infty. \end{cases}$$

Proof of the triangle inequality of the norm $\|\cdot\|_{\ell^p}$ is the content of Exercise 1.5.11. □

Example 1.2.5 (a) The standard norm for $C[a, b]$ is the *maximum norm*

$$\|f\|_\infty = \max_{a\leq x\leq b} |f(x)|, \quad f \in C[a, b].$$

This is also the norm for $C_p(2\pi)$ (with $a = 0$ and $b = 2\pi$), the space of continuous 2π-periodic functions introduced in Example 1.1.2 (f).

(b) For an integer $k > 0$, the standard norm for $C^k[a, b]$ is

$$\|f\|_{k,\infty} = \max_{0\leq j\leq k} \|f^{(j)}\|_\infty, \quad f \in C^k[a, b].$$

This is also the standard norm for $C^k_p(2\pi)$. □

With the notion of a norm for V we can introduce a topology for V, and speak about *open* and *closed* sets in V.

Definition 1.2.6 *Let $(V, \|\cdot\|)$ be a normed space. Given $v_0 \in V$ and $r > 0$, the sets*

$$B(v_0, r) = \{v \in V \mid \|v - v_0\| < r\},$$
$$\overline{B}(v_0, r) = \{v \in V \mid \|v - v_0\| \leq r\}$$

are called the open and closed balls centered at v_0 with radius r. When $r = 1$ and $v_0 = 0$, we have unit balls.

Definition 1.2.7 *Let $A \subset V$ be a set in a normed linear space. The set A is* open *if for every $v \in A$, there is an $r > 0$ such that $B(v, r) \subset A$. The set A is* closed *in V if its complement $V \backslash A$ is open in V.*

1.2.1 Convergence

With the notion of a norm at our disposal, we can define the important concept of convergence.

Definition 1.2.8 *Let V be a normed space with the norm $\|\cdot\|$. A sequence $\{u_n\} \subset V$ is* convergent *to $u \in V$ if*

$$\lim_{n\to\infty} \|u_n - u\| = 0.$$

We say that u is the limit of the sequence $\{u_n\}$, and write $u_n \to u$ as $n \to \infty$, or $\lim_{n\to\infty} u_n = u$.

It can be verified that any sequence can have at most one limit.

Definition 1.2.9 *A function $f : V \to \mathbb{R}$ is said to be* continuous *at $u \in V$ if for any sequence $\{u_n\}$ with $u_n \to u$, we have $f(u_n) \to f(u)$ as $n \to \infty$. The function f is said to be continuous on V if it is continuous at every $u \in V$.*

Proposition 1.2.10 *The norm function $\|\cdot\|$ is continuous.*

Proof. We need to show that if $u_n \to u$, then $\|u_n\| \to \|u\|$. This follows from the *backward triangle inequality* (Exercise 1.2.1)

$$|\, \|u\| - \|v\| \,| \leq \|u - v\| \quad \forall\, u, v \in V, \tag{1.2.5}$$

derived from the triangle inequality. □

Example 1.2.11 Consider the space $V = C[0,1]$. Let $x_0 \in [0,1]$. We define the function

$$\ell_{x_0}(v) = v(x_0), \quad v \in V.$$

Assume $v_n \to v$ in V as $n \to \infty$. Then

$$|\ell_{x_0}(v_n) - \ell_{x_0}(v)| \le \|v_n - v\|_V \to 0 \quad \text{as } n \to \infty.$$

Hence, the point value function ℓ_{x_0} is continuous on $C[0,1]$. □

We have seen that on a linear space various norms can be defined. Different norms give different measures of size for a given vector in the space. Consequently, different norms may give rise to different forms of convergence.

Definition 1.2.12 *We say two norms $\|\cdot\|_{(1)}$ and $\|\cdot\|_{(2)}$ are equivalent if there exist positive constants c_1, c_2 such that*

$$c_1\|v\|_{(1)} \le \|v\|_{(2)} \le c_2\|v\|_{(1)} \quad \forall\, v \in V.$$

With two such equivalent norms, a sequence $\{u_n\}$ converges in one norm if and only if it converges in the other norm:

$$\lim_{n\to\infty} \|u_n - u\|_{(1)} = 0 \quad \Longleftrightarrow \quad \lim_{n\to\infty} \|u_n - u\|_{(2)} = 0.$$

Conversely, if each sequence converging with respect to one norm also converges with respect to the other norm, then the two norms are equivalent; proof of this statement is left as Exercise 1.2.15.

Example 1.2.13 For the norms (1.2.2)–(1.2.3) on $\mathbb{R}^d$, it is straightforward to show

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_p \le d^{1/p}\|\boldsymbol{x}\|_\infty \quad \forall\, \boldsymbol{x} \in \mathbb{R}^d. \tag{1.2.6}$$

So all the norms $\|\boldsymbol{x}\|_p$, $1 \le p \le \infty$, on $\mathbb{R}^d$ are equivalent. □

More generally, we have the following well-known result. For a proof, see [15, p. 483].

Theorem 1.2.14 *Over a finite dimensional space, any two norms are equivalent.*

Thus, on a finite dimensional space, different norms lead to the same convergence notion. Over an infinite dimensional space, however, such a statement is no longer valid.

Example 1.2.15 Let V be the space of all continuous functions on $[0,1]$. For $u \in V$, in analogy with Example 1.2.3, we may define the following norms

$$\|v\|_p = \left[\int_0^1 |v(x)|^p dx\right]^{1/p}, \quad 1 \le p < \infty, \tag{1.2.7}$$

$$\|v\|_\infty = \sup_{0\le x\le 1} |v(x)|. \tag{1.2.8}$$

Now consider a sequence of functions $\{u_n\} \subset V$, defined by

$$u_n(x) = \begin{cases} 1 - nx, & 0 \le x \le \dfrac{1}{n}, \\ 0, & \dfrac{1}{n} < x \le 1. \end{cases}$$

It is easy to show that

$$\|u_n\|_p = [n(p+1)]^{-1/p}, \quad 1 \le p < \infty.$$

Thus we see that the sequence $\{u_n\}$ converges to $u = 0$ in the norm $\|\cdot\|_p$, $1 \le p < \infty$. On the other hand,

$$\|u_n\|_\infty = 1, \quad n \ge 1,$$

so $\{u_n\}$ does not converge to $u = 0$ in the norm $\|\cdot\|_\infty$. □

As we have seen in the last example, in an infinite dimensional space, some norms are not equivalent. Convergence defined by one norm can be stronger than that by another.

Example 1.2.16 Consider again the space of all continuous functions on $[0,1]$, and the family of norms $\|\cdot\|_p$, $1 \le p < \infty$, and $\|\cdot\|_\infty$. We have, for any $p \in [1,\infty)$,

$$\|v\|_p \le \|v\|_\infty \quad \forall\, v \in V.$$

Therefore, convergence in $\|\cdot\|_\infty$ implies convergence in $\|\cdot\|_p$, $1 \le p < \infty$, but not conversely (see Example 1.2.15). Convergence in $\|\cdot\|_\infty$ is usually called *uniform convergence*. □

With the notion of convergence, we can define the concept of an infinite series in a normed space.

Definition 1.2.17 *Let $\{v_n\}_{n=1}^\infty$ be a sequence in a normed space V. Define the partial sums $s_n = \sum_{i=1}^n v_i$, $n = 1, 2, \cdots$. If $s_n \to s$ in V, then we say the series $\sum_{i=1}^\infty v_i$ converges, and write*

$$\sum_{i=1}^\infty v_i = \lim_{n\to\infty} s_n = s.$$

Definition 1.2.18 *Let $V_1 \subset V_2$ be two subsets in a normed space V. We say the set V_1 is dense in V_2 if for any $u \in V_2$ and any $\varepsilon > 0$, there is a $v \in V_1$ such that $\|v - u\| < \varepsilon$.*

Example 1.2.19 Let $p \in [1, \infty)$ and $\Omega \subset \mathbb{R}^d$ be an open bounded set. Then the subspace $C_0^\infty(\Omega)$ is dense in $L^p(\Omega)$. The subspace of all the polynomials is also dense in $L^p(\Omega)$. □

We now extend the definition of a basis to an infinite dimensional normed space.

Definition 1.2.20 *Suppose V is an infinite dimensional normed space.*
(a) *We say that V has a countably-infinite basis if there is a sequence $\{v_i\}_{i\geq 1} \subset V$ for which the following is valid: For each $v \in V$, we can find scalars $\{\alpha_{n,i}\}_{i=1}^n$, $n = 1, 2, \ldots$, such that*

$$\left\| v - \sum_{i=1}^{n} \alpha_{n,i} v_i \right\| \to 0 \quad \text{as } n \to \infty.$$

The space V is also said to be separable. *The sequence $\{v_i\}_{i\geq 1}$ is called a basis if any finite subset of the sequence is linearly independent.*
(b) *We say that V has a* Schauder basis *$\{v_n\}_{n\geq 1}$ if for each $v \in V$, it is possible to write*

$$v = \sum_{n=1}^{\infty} \alpha_n v_n$$

as a convergent series in V for a unique choice of scalars $\{\alpha_n\}_{n\geq 1}$.

We see that the normed space V is separable if it has a countable dense subset. From Example 1.2.19, we conclude that for $p \in [1, \infty)$, $L^p(\Omega)$ is separable since the set of all the polynomials with rational coefficients is countable and is dense in $L^p(\Omega)$.

From the uniqueness requirement for a Schauder basis, we deduce that $\{v_n\}$ must be independent. A normed space having a Schauder basis can be shown to be separable. However, the converse is not true; see [77] for an example of a separable Banach space that does not have a Schauder basis. In the space ℓ^2, $\{\boldsymbol{e}_j = (0, \cdots, 0, 1_j, 0, \cdots)\}_{j=1}^\infty$ forms a Schauder basis since any $\boldsymbol{x} = (x_1, x_2, \cdots) \in \ell^2$ can be uniquely written as $\boldsymbol{x} = \sum_{j=1}^\infty x_j \boldsymbol{e}_j$. It can be proved that the set $\{1, \cos nx, \sin nx\}_{n=1}^\infty$ forms a Schauder basis in $L^p(-\pi, \pi)$ for $p \in (1, \infty)$; see the discussion in Section 4.1.

1.2.2 Banach spaces

The concept of a normed space is usually too general, and special attention is given to a particular type of normed space called a *Banach space*.

Definition 1.2.21 *Let V be a normed space. A sequence $\{u_n\} \subset V$ is called a* Cauchy sequence *if*

$$\lim_{m,n\to\infty} \|u_m - u_n\| = 0.$$

Obviously, a convergent sequence is a Cauchy sequence. In other words, being a Cauchy sequence is a necessary condition for a sequence to converge. Note that in showing convergence with Definition 1.2.8, one has to know the limit, and this is not convenient in many circumstances. On the contrary, it is usually relatively easier to determine if a given sequence is a Cauchy sequence. So it is natural to ask if a Cauchy sequence is convergent. In the finite dimensional space $\mathbb{R}^d$, any Cauchy sequence is convergent. However, in a general infinite dimensional space, a Cauchy sequence may fail to converge, as is demonstrated in the next example.

Example 1.2.22 Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. For $v \in C(\overline{\Omega})$ and $1 \le p < \infty$, define the p-norm

$$\|v\|_p = \left[\int_\Omega |v(\boldsymbol{x})|^p dx\right]^{1/p}. \tag{1.2.9}$$

Here, $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ and $dx = dx_1 dx_2 \cdots dx_d$. In addition, define the ∞-*norm* or *maximum norm*

$$\|v\|_\infty = \max_{\boldsymbol{x}\in\overline{\Omega}} |v(\boldsymbol{x})|.$$

The space $C(\overline{\Omega})$ with $\|\cdot\|_\infty$ is a Banach space, since the uniform limit of continuous functions is itself continuous.

The space $C(\overline{\Omega})$ with the norm $\|\cdot\|_p$, $1 \le p < \infty$, is not a Banach space. To illustrate this, we consider the space $C[0,1]$ and a sequence in $C[0,1]$ defined as follows:

$$u_n(x) = \begin{cases} 0, & 0 \le x \le 1/2 - 1/(2n), \\ n\,x - (n-1)/2, & 1/2 - 1/(2n) \le x \le 1/2 + 1/(2n), \\ 1, & 1/2 + 1/(2n) \le x \le 1. \end{cases}$$

Let

$$u(x) = \begin{cases} 0, & 0 \le x < 1/2, \\ 1, & 1/2 < x \le 1. \end{cases}$$

Then $\|u_n - u\|_p \to 0$ as $n \to \infty$, i.e., the sequence $\{u_n\}$ converges to u in the norm $\|\cdot\|_p$. But obviously no matter how we define $u(1/2)$, the limit function u is not continuous. □

Although a Cauchy sequence is not necessarily convergent, it does converge if it has a convergent subsequence.

Proposition 1.2.23 *If a Cauchy sequence contains a convergent subsequence, then the entire sequence converges to the same limit.*

Proof. Let $\{u_n\}$ be a Cauchy sequence in a normed space V, with a subsequence $\{u_{n_j}\}$ converging to $u \in V$. Then for any $\varepsilon > 0$, there exist positive integers n_0 and j_0 such that

$$\|u_m - u_n\| \leq \frac{\varepsilon}{2} \quad \forall\, m, n \geq n_0,$$
$$\|u_{n_j} - u\| \leq \frac{\varepsilon}{2} \quad \forall\, j \geq j_0.$$

Let $N = \max\{n_0, n_{j_0}\}$. Then

$$\|u_n - u\| \leq \|u_n - u_N\| + \|u_N - u\| \leq \varepsilon \quad \forall\, n \geq N.$$

Therefore, $u_n \to u$ as $n \to \infty$. □

Definition 1.2.24 *A normed space is said to be* complete *if every Cauchy sequence from the space converges to an element in the space. A complete normed space is called a* Banach space.

Example of Banach spaces include $C([a,b])$ and $L^p(a,b)$, $1 \leq p \leq \infty$, with their standard norms.

1.2.3 Completion of normed spaces

It is important to be able to deal with function spaces using a norm of our choice, as such a norm is often important or convenient in the formulation of a problem or in the analysis of a numerical method. The following theorem allows us to do this. A proof is discussed in [135, p. 84].

Theorem 1.2.25 *Let V be a normed space. Then there is a complete normed space W with the following properties:*
(a) *There is a subspace $\widehat{V} \subset W$ and a bijective (one-to-one and onto) linear function $\mathcal{I} : V \to \widehat{V}$ with*

$$\|\mathcal{I}v\|_W = \|v\|_V \quad \forall\, v \in V.$$

The function $\mathcal{I}$ is called an isometric isomorphism *of the spaces V and $\widehat{V}$.*
(b) *The subspace $\widehat{V}$ is dense in W, i.e., for any $w \in W$, there is a sequence $\{\widehat{v}_n\} \subset \widehat{V}$ such that*

$$\|w - \widehat{v}_n\|_W \to 0 \quad \text{as } n \to \infty.$$

The space W is called the completion *of V, and W is unique up to an isometric isomorphism.*

The spaces V and $\widehat{V}$ are generally identified, meaning no distinction is made between them. However, we also consider cases where it is important to note the distinction. An important example of the theorem is to let V be the rational numbers and W be the real numbers $\mathbb{R}$. One way in which $\mathbb{R}$ can be defined is as a set of equivalence classes of Cauchy sequences of rational numbers, and $\widehat{V}$ can be identified with those equivalence classes of Cauchy sequences whose limit is a rational number. A proof of the above theorem can be made by mimicking this commonly used construction of the real numbers from the rational numbers.

Theorem 1.2.25 guarantees the existence of a unique abstract completion of an arbitrary normed vector space. However, it is often possible, and indeed desirable, to give a more concrete definition of the completion of a given normed space; much of the subject of *real analysis* is concerned with this topic. In particular, the subject of *Lebesgue measure and integration* deals with the completion of $C(\overline{\Omega})$ under the norms of (1.2.9), $\|\cdot\|_p$ for $1 \leq p < \infty$. A complete development of Lebesgue measure and integration is given in any standard textbook on real analysis; for example, see Royden [198] or Rudin [199]. We do not introduce formally and rigorously the concepts of *measurable set* and *measurable function*. Rather we think of measure theory intuitively as described in the following paragraphs. Our rationale for this is that the details of Lebesgue measure and integration can often be bypassed in most of the material we present in this text.

Measurable subsets of $\mathbb{R}$ include the standard open and closed intervals with which we are familiar. Multi-variable extensions of intervals to $\mathbb{R}^d$ are also measurable, together with countable unions and intersections of them. In particular, open sets and closed sets are measurable. Intuitively, the measure of a set $D \subset \mathbb{R}^d$ is its "length", "area", "volume", or suitable generalization; and we denote the *measure* of D by meas(D). For a formal discussion of measurable set, see Royden [198] or Rudin [199].

To introduce the concept of measurable function, we begin by defining a step function. A function v on a measurable set D is a *step function* if D can be decomposed into a finite number of pairwise disjoint measurable subsets $D_1, \ldots, D_k$ with $v(\boldsymbol{x})$ constant over each D_j. We say a function v on D is a *measurable function* if it is the pointwise limit of a sequence of step functions. This includes, for example, all continuous functions on D.

For each such measurable set D_j, we define a *characteristic function*

$$\chi_j(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in D_j, \\ 0, & \boldsymbol{x} \notin D_j. \end{cases}$$

A general step function over the decomposition $D_1, \ldots, D_k$ of D can then be written as

$$v(\boldsymbol{x}) = \sum_{j=1}^{k} \alpha_j \chi_j(\boldsymbol{x}), \quad \boldsymbol{x} \in D \tag{1.2.10}$$

with $\alpha_1, \ldots, \alpha_k$ scalars. For a general measurable function v over D, we write it as a limit of step functions v_k over D:

$$v(\boldsymbol{x}) = \lim_{k\to\infty} v_k(\boldsymbol{x}), \quad \boldsymbol{x} \in D. \tag{1.2.11}$$

We say *two measurable functions are equal almost everywhere* if the set of points on which they differ is a set of measure zero. For notation, we write

$$v = w \text{ (a.e.)}$$

to indicate that v and w are equal almost everywhere. Given a measurable function v on D, we introduce the concept of an equivalence class of equivalent functions:

$$[v] = \{w \mid w \text{ measurable on } D \text{ and } v = w \text{ (a.e.)}\}.$$

For most purposes, we generally consider elements of an equivalence class $[v]$ as being a single function v.

We define the Lebesgue integral of a step function v over D, given in (1.2.10), by

$$\int_D v(\boldsymbol{x})\, dx = \sum_{j=1}^{k} \alpha_j \,\text{meas}(D_j).$$

For a general measurable function, given in (1.2.11), define the Lebesgue integral of v over D by

$$\int_D v(\boldsymbol{x})\, dx = \lim_{k\to\infty} \int_D v_k(\boldsymbol{x})\, dx.$$

Note that the Lebesgue integrals of elements of an equivalence class $[v]$ are identical. There are a great many properties of Lebesgue integration, and we refer the reader to any text on real analysis for further details. Here we only record two important theorems for later referral.

Theorem 1.2.26 (Lebesgue Dominated Convergence Theorem) *Suppose $\{f_n\}$ is a sequence of Lebesgue integrable functions converging a.e. to f on a measurable set D. If there exists a Lebesgue integrable function g such that*

$$|f_n(\boldsymbol{x})| \le g(\boldsymbol{x}) \quad \text{a.e. in } D,\ n \ge 1,$$

then the limit f is Lebesgue integrable and

$$\lim_{n\to\infty} \int_D f_n(\boldsymbol{x})\, dx = \int_D f(\boldsymbol{x})\, dx.$$

Theorem 1.2.27 (Fubini's Theorem) *Assume $D_1 \subset \mathbb{R}^{d_1}$ and $D_2 \subset \mathbb{R}^{d_2}$ are Lebesgue measurable sets, and let f be a Lebesgue integrable function*

on $D = D_1 \times D_2$. Then for a.e. $\boldsymbol{x} \in D_1$, the function $f(\boldsymbol{x}, \cdot)$ is Lebesgue integrable on D_2, $\int_{D_2} f(\boldsymbol{x}, \boldsymbol{y})\, dy$ is integrable on D_1, and

$$\int_{D_1} \left[\int_{D_2} f(\boldsymbol{x}, \boldsymbol{y})\, dy \right] dx = \int_D f(\boldsymbol{x}, \boldsymbol{y})\, dx\, dy.$$

Similarly, for a.e. $\boldsymbol{y} \in D_2$, the function $f(\cdot, \boldsymbol{y})$ is Lebesgue integrable on D_1, $\int_{D_1} f(\boldsymbol{x}, \boldsymbol{y})\, dx$ is integrable on D_2, and

$$\int_{D_2} \left[\int_{D_1} f(\boldsymbol{x}, \boldsymbol{y})\, dx \right] dy = \int_D f(\boldsymbol{x}, \boldsymbol{y})\, dx\, dy.$$

Let Ω be an open set in $\mathbb{R}^d$. For $1 \le p < \infty$, introduce

$$L^p(\Omega) = \{[v] \mid v \text{ measurable on } \Omega \text{ and } \|v\|_p < \infty\}.$$

The norm $\|v\|_p$ is defined as in (1.2.9), although now we use Lebesgue integration rather than Riemann integration. For v measurable on Ω, denote

$$\|v\|_\infty = \operatorname*{ess\,sup}_{\boldsymbol{x} \in \Omega} |v(\boldsymbol{x})| \equiv \inf_{\text{meas}(\Omega')=0} \sup_{\boldsymbol{x} \in \Omega \setminus \Omega'} |v(\boldsymbol{x})|,$$

where "meas$(\Omega') = 0$" means Ω' is a measurable set with measure zero. Then we define

$$L^\infty(\Omega) = \{[v] \mid v \text{ measurable on } \Omega \text{ and } \|v\|_\infty < \infty\}.$$

The spaces $L^p(\Omega)$, $1 \le p < \infty$, are Banach spaces, and they are concrete realizations of the abstract completion of $C(\overline{\Omega})$ under the norm of (1.2.9). The space $L^\infty(\Omega)$ is also a Banach space, but it is much larger than the space $C(\overline{\Omega})$ with the ∞-norm $\|\cdot\|_\infty$. Additional discussion of the spaces $L^p(\Omega)$ is given in Section 1.5.

More generally, let w be a positive continuous function on Ω, called a *weight function*. We can define weighted spaces $L^p_w(\Omega)$ as follows

$$L^p_w(\Omega) = \left\{ v \text{ measurable} \;\middle|\; \int_\Omega w(\boldsymbol{x})\, |v(\boldsymbol{x})|^p\, dx < \infty \right\}, \quad p \in [1, \infty),$$
$$L^\infty_w(\Omega) = \{v \text{ measurable} \mid \operatorname{ess\,sup}_\Omega w(\boldsymbol{x})\, |v(\boldsymbol{x})| < \infty\}.$$

These are Banach spaces with the norms

$$\|v\|_{p,w} = \left[\int_\Omega w(\boldsymbol{x})\, |v(\boldsymbol{x})|^p\, dx \right]^{1/p}, \quad p \in [1, \infty),$$
$$\|v\|_{\infty,w} = \operatorname*{ess\,sup}_{\boldsymbol{x} \in \Omega} w(\boldsymbol{x}) |v(\boldsymbol{x})|.$$

The space $C(\overline{\Omega})$ of Example 1.1.2 (c) with the norm

$$\|v\|_{C(\overline{\Omega})} = \max_{\boldsymbol{x} \in \overline{\Omega}} |v(\boldsymbol{x})|$$

is also a Banach space, and it can be considered as a proper subset of $L^\infty(\Omega)$. See Example 2.5.3 for a situation where it is necessary to distinguish between $C(\overline{\Omega})$ and the subspace of $L^\infty(\Omega)$ to which it is isometric and isomorphic.

Example 1.2.28 (a) For any integer $m \geq 0$, the normed spaces $C^m[a,b]$ and $C_p^k(2\pi)$ of Example 1.2.5 (b) are Banach spaces.
(b) Let $1 \leq p < \infty$. As an alternative norm on $C^m[a,b]$, introduce

$$\|f\| = \left[\sum_{j=0}^{m} \|f^{(j)}\|_p^p\right]^{1/p}.$$

The space $C^m[a,b]$ is not complete with this norm. Its completion is denoted by $W^{m,p}(a,b)$, an example of a *Sobolev space*. It can be shown that if $f \in W^{m,p}(a,b)$, then $f, f', \ldots, f^{(m-1)}$ are continuous, and $f^{(m)}$ exists almost everywhere and belongs to $L^p(a,b)$. This Sobolev space and its multi-variable generalizations are discussed at length in Chapter 7. □

A knowledge of the theory of Lebesgue measure and integration is very helpful in dealing with problems defined on spaces of Lebesgue integrable functions. Nonetheless, many results can be proven by referring to only the original space and its associated norm, say $C(\overline{\Omega})$ with $\|\cdot\|_p$, from which a Banach space is obtained by a completion argument, say $L^p(\Omega)$. We return to this in Theorem 2.4.1 of Chapter 2.

Exercise 1.2.1 Prove the backward triangle inequality (1.2.5). More generally,

$$|\|u\| - \|v\|| \leq \|u \pm v\| \leq \|u\| + \|v\| \quad \forall\, u, v \in V.$$

Exercise 1.2.2 Let V be a normed space. Show that the vector addition and scalar multiplication are continuous operations, i.e., from $u_n \to u$, $v_n \to v$ and $\alpha_n \to \alpha$, we can conclude that

$$u_n + v_n \to u + v, \quad \alpha_n v_n \to \alpha\, v.$$

Exercise 1.2.3 Show that $\|\cdot\|_\infty$ is a norm on $C(\overline{\Omega})$, with Ω a bounded open set in $\mathbb{R}^d$.

Exercise 1.2.4 Show that $\|\cdot\|_\infty$ is a norm on $L^\infty(\Omega)$, with Ω a bounded open set in $\mathbb{R}^d$.

Exercise 1.2.5 Show that $\|\cdot\|_1$ is a norm on $L^1(\Omega)$, with Ω a bounded open set in $\mathbb{R}^d$.

Exercise 1.2.6 Show that for $p = 1, 2, \infty$, $\|\cdot\|_p$ defined by (1.2.2)–(1.2.3) is a norm in the space $\mathbb{R}^d$.

Exercise 1.2.7 Show that the norm $\|\cdot\|_p$ defined by (1.2.2)–(1.2.3) has a monotonicity property with respect to p:

$$1 \le p \le q \le \infty \quad \Longrightarrow \quad \|\boldsymbol{x}\|_q \le \|\boldsymbol{x}\|_p \quad \forall\, \boldsymbol{x} \in \mathbb{R}^d.$$

Exercise 1.2.8 Define $C^\alpha[a,b]$, $0 < \alpha \le 1$, as the set of all $f \in C[a,b]$ for which

$$M_\alpha(f) \equiv \sup_{\substack{a \le x,y \le b \\ x \ne y}} \frac{|f(x) - f(y)|}{|x-y|^\alpha} < \infty.$$

Define $\|f\|_\alpha = \|f\|_\infty + M_\alpha(f)$. Show $C^\alpha[a,b]$ with this norm is complete.

Exercise 1.2.9 Define $C_b[0,\infty)$ as the set of all functions f that are continuous on $[0,\infty)$ and satisfy

$$\|f\|_\infty \equiv \sup_{x \ge 0} |f(x)| < \infty.$$

Show $C_b[0,\infty)$ with this norm is complete.

Exercise 1.2.10 Does the formula (1.2.2) define a norm on $\mathbb{R}^d$ for $0 < p < 1$?

Exercise 1.2.11 Consider the norm (1.2.7) on $V = C[0,1]$. For $1 \le p < q < \infty$, construct a sequence $\{v_n\} \subset C[0,1]$ such that as $n \to \infty$, $\|v_n\|_p \to 0$ and $\|v_n\|_q \to \infty$.

Exercise 1.2.12 Prove the equivalence of the following norms on $C^1[0,1]$:

$$\|f\|_a \equiv |f(0)| + \int_0^1 |f'(x)|\,dx,$$
$$\|f\|_b \equiv \int_0^1 |f(x)|\,dx + \int_0^1 |f'(x)|\,dx.$$

Hint: Recall the integral mean value theorem: Given $g \in C[0,1]$, there is a $\xi \in [0,1]$ such that

$$\int_0^1 g(x)\,dx = g(\xi).$$

Exercise 1.2.13 Let V_1 and V_2 be normed spaces with norms $\|\cdot\|_1$ and $\|\cdot\|_2$. Recall that the product space $V_1 \times V_2$ is defined by

$$V_1 \times V_2 = \{(v_1, v_2) \mid v_1 \in V_1,\, v_2 \in V_2\}.$$

Show that the quantities $\max\{\|v_1\|_1, \|v_2\|_2\}$ and $(\|v_1\|_1^p + \|v_2\|_2^p)^{1/p}$, $1 \le p < \infty$ all define norms on the space $V_1 \times V_2$.

Exercise 1.2.14 Over the space $C^1[0,1]$, determine which of the following is a norm, and which is only a semi-norm:

(a) $\max\limits_{0 \le x \le 1} |u(x)|$;

(b) $\max\limits_{0 \le x \le 1} [|u(x)| + |u'(x)|]$;

(c) $\max\limits_{0\le x\le 1} |u'(x)|$;

(d) $|u(0)| + \max\limits_{0\le x\le 1} |u'(x)|$;

(e) $\max\limits_{0\le x\le 1} |u'(x)| + \int_{0.1}^{0.2} |u(x)|\,dx$.

Exercise 1.2.15 Show that two norms on a linear space are equivalent if and only if each sequence converging with respect to one norm also converges with respect to the other norm.

Exercise 1.2.16 Over a normed space $(V, \|\cdot\|)$, we define a function of two variables $d(u,v) = \|u - v\|$. Show that $d(\cdot,\cdot)$ is a *distance function*, in other words, $d(\cdot,\cdot)$ has the following properties of an ordinary distance between two points:

(a) $d(u,v) \ge 0$ for any $u, v \in V$, and $d(u,v) = 0$ if and only if $u = v$;

(b) $d(u,v) = d(v,u)$ for any $u, v \in V$;

(c) (the triangle inequality) $d(u,w) \le d(u,v) + d(v,w)$ for any $u, v, w \in V$.

Also show that the non-negativity of $d(\cdot,\cdot)$ can be deduced from the property "$d(u,v) = 0$ if and only if $u = v$" together with (b) and (c).

A linear space endowed with a distance function is called a *metric space*. Certainly a normed space can be viewed as a metric space. There are examples of metrics (distance functions) which are not generated by any norm, though.

Exercise 1.2.17 Show that in a Banach space, if $\{v_n\}_{n=1}^{\infty}$ is a sequence satisfying $\sum_{n=1}^{\infty} \|v_n\| < \infty$, then the series $\sum_{n=1}^{\infty} v_n$ converges. Such a series is said to converge absolutely.

Exercise 1.2.18 Let V be a Banach space, and $\lambda \in (0,2)$. Starting with any two points $v_0, v_1 \in V$, define a sequence $\{v_n\}_{n=0}^{\infty}$ by the formula

$$v_{n+1} = \lambda\, v_n + (1-\lambda)\, v_{n-1}, \quad n \ge 1.$$

Show that the sequence $\{v_n\}_{n=0}^{\infty}$ converges.

Exercise 1.2.19 Let V be a normed space, $V_0 \subset V$ a closed subspace. The quotient space V/V_0 is defined to be the space of all the classes

$$[v] = \{v + v_0 \mid v_0 \in V_0\}.$$

Prove that the formula

$$\|[v]\|_{V/V_0} = \inf_{v_0 \in V_0} \|v + v_0\|_V$$

defines a norm on V/V_0. Show that if V is a Banach space, then V/V_0 is a Banach space.

Exercise 1.2.20 Assuming a knowledge of Lebesgue integration, show that

$$W^{1,2}(a,b) \subset C[a,b].$$

Generalize this result to the space $W^{m,p}(a,b)$ with other values of m and p. *Hint*: For $v \in W^{1,2}(a,b)$, use

$$v(x) - v(y) = \int_x^y v'(z)\, dz\,.$$

Exercise 1.2.21 On $C^1[0,1]$, define

$$(u,v)_* = u(0)\, v(0) + \int_0^1 u'(x)\, v'(x)\, dx$$

and

$$\|v\|_* = \sqrt{(v,v)_*}\,.$$

Show that

$$\|v\|_\infty \le c\, \|v\|_* \quad \forall\, v \in C^1[0,1]$$

for a suitably chosen constant c.

Exercise 1.2.22 Apply Theorem 1.2.26 to show the following form of Lebesgue Dominated Convergence Theorem: Suppose a sequence $\{f_n\} \subset L^p(D)$, $1 \le p < \infty$, converges a.e. to f on a measurable set D. If there exists a function $g \in L^p(D)$ such that

$$|f_n(\boldsymbol{x})| \le g(\boldsymbol{x}) \quad \text{a.e. in } D,\ n \ge 1,$$

then the limit $f \in L^p(D)$ and

$$\lim_{n\to\infty} \|f_n - f\|_{L^p(D)} = 0.$$

1.3 Inner product spaces

In studying linear problems, inner product spaces are usually used. These are the spaces where a norm can be defined through the inner product and the notion of orthogonality of two elements can be introduced. The inner product in a general space is a generalization of the usual scalar product (or dot product) in the plane $\mathbb{R}^2$ or the space $\mathbb{R}^3$.

Definition 1.3.1 *Let V be a linear space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. An* inner product $(\cdot,\cdot)$ *is a function from $V \times V$ to $\mathbb{K}$ with the following properties.*

1. *For any $v \in V$, $(v,v) \ge 0$ and $(v,v) = 0$ if and only if $v = 0$.*
2. *For any $u, v \in V$, $(u,v) = \overline{(v,u)}$.*
3. *For any $u, v, w \in V$, any $\alpha, \beta \in \mathbb{K}$, $(\alpha\, u + \beta\, v, w) = \alpha\, (u,w) + \beta\, (v,w)$.*

The space V together with the inner product $(\cdot,\cdot)$ is called an inner product space. *When the definition of the inner product $(\cdot,\cdot)$ is clear from the context, we simply say V is an inner product space. When $\mathbb{K}=\mathbb{R}$, V is called a* real inner product space*, whereas if $\mathbb{K}=\mathbb{C}$, V is a* complex inner product space*.*

In the case of a real inner product space, the second axiom reduces to the symmetry of the inner product:

$$(u,v)=(v,u)\quad \forall\, u,v\in V.$$

For an inner product, there is an important property called the Schwarz inequality.

Theorem 1.3.2 (Schwarz inequality) *If V is an inner product space, then*

$$|(u,v)|\le\sqrt{(u,u)\,(v,v)}\qquad \forall\, u,v\in V,$$

and the equality holds if and only if u and v are linearly dependent.

Proof. We give the proof only for the real case; the complex case is treated in Exercise 1.3.2. The result is obviously true if either $u=0$ or $v=0$. Now suppose $u\neq 0$, $v\neq 0$. Define

$$\phi(t)=(u+t\,v,u+t\,v)=(u,u)+2\,(u,v)\,t+(v,v)\,t^2,\quad t\in\mathbb{R}.$$

The function ϕ is quadratic and non-negative, so its discriminant must be non-positive,

$$[2\,(u,v)]^2-4\,(u,u)\,(v,v)\le 0,$$

i.e., the Schwarz inequality is valid. For $v\neq 0$, the equality holds if and only if $u=-t\,v$ for some $t\in\mathbb{R}$.

See Exercise 1.3.1 for another proof. □

An inner product $(\cdot,\cdot)$ induces a norm through the formula

$$\|v\|=\sqrt{(v,v)},\quad v\in V.$$

In verifying the triangle inequality for the quantity thus defined, we need to use the above Schwarz inequality. Moreover, equality

$$\|u+v\|=\|u\|+\|v\|$$

holds if and only if u or v is a non-negative multiple of the other. Proof of these statement is left as Exercise 1.3.4.

Proposition 1.3.3 *An inner product is continuous with respect to its induced norm. In other words, if $\|\cdot\|$ is the norm defined by $\|v\|=\sqrt{(v,v)}$, then $\|u_n-u\|\to 0$ and $\|v_n-v\|\to 0$ as $n\to\infty$ imply*

$$(u_n,v_n)\to(u,v)\quad \text{as } n\to\infty.$$

In particular, if $u_n \to u$, then for any v,

$$(u_n, v) \to (u, v) \quad \text{as } n \to \infty.$$

Proof. Since $\{u_n\}$ and $\{v_n\}$ are convergent, they are bounded, i.e., for some $M < \infty$, $\|u_n\| \le M$, $\|v_n\| \le M$ for any n. We write

$$(u_n, v_n) - (u, v) = (u_n - u, v_n) + (u, v_n - v).$$

Using the Schwarz inequality, we have

$$\begin{aligned} |(u_n, v_n) - (u, v)| &\le \|u_n - u\| \, \|v_n\| + \|u\| \, \|v_n - v\| \\ &\le M \, \|u_n - u\| + \|u\| \, \|v_n - v\|. \end{aligned}$$

Hence the result holds. □

Commonly seen inner product spaces are usually associated with their canonical inner products. As an example, the canonical inner product for the space $\mathbb{R}^d$ is

$$(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} x_i y_i = \boldsymbol{y}^T \boldsymbol{x}, \quad \forall\, \boldsymbol{x} = (x_1, \ldots, x_d)^T,\ \boldsymbol{y} = (y_1, \ldots, y_d)^T \in \mathbb{R}^d.$$

This inner product induces the Euclidean norm

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} = \left(\sum_{i=1}^{d} |x_i|^2 \right)^{1/2}.$$

When we talk about the space $\mathbb{R}^d$, implicitly we understand the inner product and the norm are the ones defined above, unless stated otherwise. For the complex space $\mathbb{C}^d$, the inner product and the corresponding norm are

$$(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} x_i \overline{y_i} = \boldsymbol{y}^* \boldsymbol{x}, \quad \forall\, \boldsymbol{x} = (x_1, \ldots, x_d)^T,\ \boldsymbol{y} = (y_1, \ldots, y_d)^T \in \mathbb{C}^d$$

and

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})} = \left(\sum_{i=1}^{d} |x_i|^2 \right)^{1/2}.$$

The space $L^2(\Omega)$ is an inner product space with the canonical inner product

$$(u, v) = \int_{\Omega} u(\boldsymbol{x}) \, \overline{v(\boldsymbol{x})} \, dx.$$

This inner product induces the standard $L^2(\Omega)$-norm

$$\|v\|_2 = \sqrt{(v, v)} = \left[\int_{\Omega} |v(\boldsymbol{x})|^2 dx \right]^{1/2}.$$

We have seen that an inner product induces a norm, which is always the norm we use on the inner product space unless stated otherwise. It is easy to show that on a complex inner product space,

$$(u,v) = \frac{1}{4}\left(\|u+v\|^2 - \|u-v\|^2 + i\|u+iv\|^2 - i\|u-iv\|^2\right), \qquad (1.3.1)$$

and on a real inner product space,

$$(u,v) = \frac{1}{4}\left(\|u+v\|^2 - \|u-v\|^2\right). \qquad (1.3.2)$$

These relations are called the *polarization identities*. Thus in any normed linear space, there can exist at most one inner product which generates the norm.

On the other hand, not every norm can be defined through an inner product. We have the following characterization for any norm induced by an inner product.

Theorem 1.3.4 *A norm $\|\cdot\|$ on a linear space V is induced by an inner product if and only if it satisfies the Parallelogram Law:*

$$\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2 \qquad \forall\, u, v \in V. \qquad (1.3.3)$$

Proof. We prove the result for the case of a real space only. Assume $\|\cdot\| = \sqrt{(\cdot,\cdot)}$ for some inner product $(\cdot,\cdot)$. Then for any $u, v \in V$,

$$\begin{aligned}\|u+v\|^2 + \|u-v\|^2 &= (u+v, u+v) + (u-v, u-v)\\ &= \left[\|u\|^2 + 2(u,v) + \|v\|^2\right]\\ &\quad + \left[\|u\|^2 - 2(u,v) + \|v\|^2\right]\\ &= 2\|u\|^2 + 2\|v\|^2.\end{aligned}$$

Conversely, assume the norm $\|\cdot\|$ satisfies the Parallelogram Law. For $u, v \in V$, let us define

$$(u,v) = \frac{1}{4}\left(\|u+v\|^2 - \|u-v\|^2\right)$$

and show that it is an inner product. First,

$$(v,v) = \frac{1}{4}\|2v\|^2 = \|v\|^2 \geq 0$$

and $(v,v) = 0$ if and only if $v = 0$. Second,

$$(u,v) = \frac{1}{4}\left(\|v+u\|^2 - \|v-u\|^2\right) = (v,u).$$

Finally, we show the linearity, which is equivalent to the following two relations:

$$(u+v, w) = (u,w) + (v,w) \quad \forall\, u, v, w \in V$$

and

$$(\alpha\, u, v) = \alpha\, (u, v) \quad \forall\, u, v \in V,\ \alpha \in \mathbb{R}\,.$$

We have

$$\begin{aligned}
(u, w) + (v, w) &= \frac{1}{4}\left(\|u + w\|^2 - \|u - w\|^2 + \|v + w\|^2 - \|v - w\|^2\right) \\
&= \frac{1}{4}\left[(\|u + w\|^2 + \|v + w\|^2) - (\|u - w\|^2 + \|v - w\|^2)\right] \\
&= \frac{1}{4}\left[\frac{1}{2}\left(\|u + v + 2\, w\|^2 + \|u - v\|^2\right)\right. \\
&\quad \left. -\frac{1}{2}\left(\|u + v - 2\, w\|^2 + \|u - v\|^2\right)\right] \\
&= \frac{1}{8}\left(\|u + v + 2\, w\|^2 - \|u + v - 2\, w\|^2\right) \\
&= \frac{1}{8}\left[2\left(\|u + v + w\|^2 + \|w\|^2\right) - \|u + v\|^2\right. \\
&\qquad \left. -2\left(\|u + v - w\|^2 + \|w\|^2\right) + \|u + v\|^2\right] \\
&= \frac{1}{4}\left(\|u + v + w\|^2 - \|u + v - w\|^2\right) \\
&= (u + v, w).
\end{aligned}$$

The proof of the second relation is more involved. For fixed $u, v \in V$, let us define a function of a real variable

$$f(\alpha) = \|\alpha\, u + v\|^2 - \|\alpha\, u - v\|^2.$$

We show that $f(\alpha)$ is a linear function of α. We have

$$\begin{aligned}
f(\alpha) - f(\beta) &= \|\alpha\, u + v\|^2 + \|\beta\, u - v\|^2 - \|\alpha\, u - v\|^2 - \|\beta\, u + v\|^2 \\
&= \frac{1}{2}\left[\|(\alpha + \beta)\, u\|^2 + \|(\alpha - \beta)\, u + 2\, v\|^2\right] \\
&\quad - \frac{1}{2}\left[\|(\alpha + \beta)\, u\|^2 + \|(\alpha - \beta)\, u - 2\, v\|^2\right] \\
&= \frac{1}{2}\left[\|(\alpha - \beta)\, u + 2\, v\|^2 - \|(\alpha - \beta)\, u - 2\, v\|^2\right] \\
&= 2\left(\left\|\frac{\alpha - \beta}{2}\, u + v\right\|^2 - \left\|\frac{\alpha - \beta}{2}\, u - v\right\|^2\right) \\
&= 2\, f\left(\frac{\alpha - \beta}{2}\right).
\end{aligned}$$

Taking $\beta = 0$ and noticing $f(0) = 0$, we find that

$$f(\alpha) = 2\, f\left(\frac{\alpha}{2}\right).$$

Thus we also have the relation

$$f(\alpha) - f(\beta) = f(\alpha - \beta).$$

From the above relations, the continuity of f, and the value $f(0) = 0$, one concludes that (see Exercise 1.3.8)

$$f(\alpha) = c_0\alpha = \alpha\, f(1) = \alpha \left(\|u+v\|^2 - \|u-v\|^2 \right)$$

from which, we get the second required relation. □

Note that if u and v form two adjacent sides of a parallelogram, then $\|u+v\|$ and $\|u-v\|$ represent the lengths of the diagonals of the parallelogram. Theorem 1.3.4 can be considered as a generalization of the Theorem of Pythagoras for right triangles.

1.3.1 Hilbert spaces

Among the inner product spaces, of particular importance are the Hilbert spaces.

Definition 1.3.5 *A complete inner product space is called a* Hilbert space.

From the definition, we see that an inner product space V is a Hilbert space if V is a Banach space under the norm induced by the inner product.

Example 1.3.6 (Some examples of Hilbert spaces)
(a) The Cartesian space $\mathbb{C}^d$ is a Hilbert space with the inner product

$$(x, y) = \sum_{i=1}^{d} x_i \overline{y_i}.$$

(b) The space $\ell^2 = \{x = \{x_i\}_{i\geq 1} \mid \sum_{i=1}^{\infty} |x_i|^2 < \infty\}$ is a linear space with

$$\alpha\, x + \beta\, y = \{\alpha\, x_i + \beta\, y_i\}_{i\geq 1}.$$

It can be shown that

$$(x, y) = \sum_{i=1}^{\infty} x_i \overline{y_i}$$

defines an inner product on ℓ^2. Furthermore, ℓ^2 becomes a Hilbert space under this inner product.
(c) The space $L^2(0, 1)$ is a Hilbert space with the inner product

$$(u, v) = \int_0^1 u(x)\, \overline{v(x)}\, dx.$$

(d) The space $L^2(\Omega)$ is a Hilbert space with the inner product

$$(u, v) = \int_\Omega u(\boldsymbol{x})\,\overline{v(\boldsymbol{x})}\,dx.$$

More generally, if $w(\boldsymbol{x})$ is a weight function on Ω, then the space

$$L^2_w(\Omega) = \left\{ v \text{ measurable} \;\middle|\; \int_\Omega |v(\boldsymbol{x})|^2 w(\boldsymbol{x})\,dx < \infty \right\}$$

is a Hilbert space with the inner product

$$(u, v)_w = \int_\Omega u(\boldsymbol{x})\,\overline{v(\boldsymbol{x})}\,w(\boldsymbol{x})\,dx.$$

This space is a weighted L^2 space. □

Example 1.3.7 Recall the Sobolev space $W^{m,p}(a, b)$ defined in Example 1.2.28. If we choose $p = 2$, then we obtain a Hilbert space. It is usually denoted by $H^m(a, b) \equiv W^{m,2}(a, b)$. The associated inner product is defined by

$$(f, g)_{H^m} = \sum_{j=0}^{m} \left(f^{(j)}, g^{(j)} \right), \quad f, g \in H^m(a, b),$$

using the standard inner product $(\cdot, \cdot)$ of $L^2(a, b)$. Recall from Exercise 1.2.20 that $H^1(a, b) \subset C[a, b]$. □

1.3.2 Orthogonality

With the notion of an inner product at our disposal, we can define the angle between two non-zero vectors u and v in a real inner product space as follows:

$$\theta = \arccos\left[\frac{(u, v)}{\|u\|\,\|v\|} \right].$$

This definition makes sense because, by the Schwarz inequality (Theorem 1.3.2), the argument of arccos is between -1 and 1. The case of a right angle is particularly important. We see that two non-zero vectors u and v form a right angle if and only if $(u, v) = 0$. In the following, we allow the inner product space V to be real or complex.

Definition 1.3.8 *Two vectors u and v are said to be* orthogonal *if $(u, v) = 0$. An element $v \in V$ is said to be orthogonal to a subset $U \subset V$, if $(u, v) = 0$ for any $u \in U$.*

By definition, the zero vector is orthogonal to any vector and any subset of the space. When some elements $u_1, \ldots, u_n$ are mutually orthogonal to each other, we have the equality

$$\|u_1 + \cdots + u_n\|^2 = \|u_1\|^2 + \cdots + \|u_n\|^2.$$

We will apply this equality for orthogonal elements repeatedly without explicitly mentioning it.

Definition 1.3.9 *Let U be a subset of an inner product space V. We define its* orthogonal complement *to be the set*

$$U^\perp = \{v \in V \mid (v, u) = 0 \ \forall\, u \in U\}.$$

As an example in $\mathbb{R}^3$, the orthogonal complement of the single element set $\{\boldsymbol{e}_1\}$ is the x_2x_3-plane $\big\{(0, x_2, x_3)^T \mid x_2, x_3 \in \mathbb{R}\big\}$.

The orthogonal complement of any set is a closed subspace (see Exercise 1.3.12).

Definition 1.3.10 *Let V be an inner product space. We say $\{v_i\}_{i\geq 1} \subset V$ forms an* orthonormal system *if*

$$(v_i, v_j) = \delta_{ij}, \quad i, j \geq 1. \tag{1.3.4}$$

If the orthonormal system is a basis of V following Definition 1.2.20, then it is called an orthonormal basis *for V.*

Sometimes we also use the notion of an *orthogonal basis*. This refers to the situation where the basis vectors are orthogonal and are not necessarily normalized to be of length 1.

Theorem 1.3.11 *Suppose $\{v_j\}_{j=1}^\infty$ is an orthonormal system in a Hilbert space V. Then we have the following conclusions.*
(a) *Bessel's inequality:*

$$\sum_{j=1}^{\infty} |(v, v_j)|^2 \leq \|v\|^2 \quad \forall\, v \in V. \tag{1.3.5}$$

(b) *For any $v \in V$, the series $\sum_{j=1}^\infty (v, v_j)\, v_j$ converges in V.*
(c) *If $v = \sum_{j=1}^\infty a_j v_j \in V$, then $a_j = (v, v_j)$.*
(d) *A series $\sum_{j=1}^\infty a_j v_j$ converges in V if and only if $\sum_{j=1}^\infty |a_j|^2 < \infty$.*

Proof. Let $v \in V$. For any positive integer n,

$$0 \leq \left\| v - \sum_{j=1}^{n} (v, v_j)\, v_j \right\|^2 = \|v\|^2 - \sum_{j=1}^{n} |(v, v_j)|^2,$$

i.e.,

$$\sum_{j=1}^{n} |(v, v_j)|^2 \le \|v\|^2.$$

Taking the limit $n \to \infty$, we obtain Bessel's inequality (1.3.5).

To prove part (b), we consider the partial sums

$$s_n = \sum_{j=1}^{n} (v, v_j)\, v_j.$$

For $m > n$, we have

$$\|s_m - s_n\|^2 = \left\| \sum_{j=n+1}^{m} (v, v_j)\, v_j \right\|^2 = \sum_{j=n+1}^{m} |(v, v_j)|^2 \to 0 \quad \text{as } n \to \infty,$$

by (a). Hence, the sequence $\{s_n\}$ is a Cauchy sequence in the Hilbert space V and is therefore convergent.

For part (c), we know from Proposition 1.3.3 that the inner product is continuous with respect to its arguments. Hence,

$$\begin{aligned}
(v, v_i) &= \left(\lim_{n\to\infty} \sum_{j=1}^{n} a_j v_j, v_i \right) \\
&= \lim_{n\to\infty} \left(\sum_{j=1}^{n} a_j v_j, v_i \right) \\
&= \lim_{n\to\infty} \sum_{j=1}^{n} a_j (v_j, v_i) \\
&= a_i.
\end{aligned}$$

Finally, for part (d), we consider the partial sums

$$s_n = \sum_{j=1}^{n} a_j v_j.$$

As in the proof of part (b), for $m > n$, we have

$$\|s_m - s_n\|^2 = \sum_{j=n+1}^{m} |a_j|^2.$$

So $\{s_n\}$ is a Cauchy sequence if and only if $\sum_{j\ge 1} |a_j|^2 < \infty$. Since V is a Hilbert space, we then get the conclusion (d). □

An orthonormal system does not need to be an orthonormal basis. The next theorem provides several ways to check if a given orthonormal system is a basis. For $\{v_j\}_{j=1}^{\infty} \subset V$ in a linear space, the symbol $\operatorname{span}\{v_j\}_{j=1}^{\infty}$ stands for the set of all the finite linear combinations of $\{v_j\}_{j=1}^{\infty}$. It is easily seen that $\operatorname{span}\{v_j\}_{j=1}^{\infty}$ is a subspace of V.

Theorem 1.3.12 *Suppose $\{v_j\}_{j=1}^{\infty}$ is an orthonormal system in a Hilbert space V. Then the following statements are equivalent.*
(a) *$\{v_j\}_{j=1}^{\infty}$ is an orthonormal basis for V.*
(b) *For any $u, v \in V$,*

$$(u, v) = \sum_{j=1}^{\infty} (u, v_j)\, \overline{(v, v_j)}. \tag{1.3.6}$$

(c) *Parseval's equality holds:*

$$\|v\|^2 = \sum_{j=1}^{\infty} |(v, v_j)|^2 \quad \forall\, v \in V. \tag{1.3.7}$$

(d) *The subspace $\operatorname{span}\{v_j\}_{j=1}^{\infty}$ is dense in V.*
(e) *For $v \in V$, if $(v, v_j) = 0$, $j \geq 1$, then $v = 0$.*

Proof. (a) $\Longrightarrow$ (b). By Theorem 1.3.11 (c) and the assumption that $\{v_j\}_{j=1}^{\infty}$ is an orthonormal basis, we can write

$$u = \sum_{j=1}^{\infty} (u, v_j)\, v_j, \quad v = \sum_{j=1}^{\infty} (v, v_j)\, v_j.$$

By the continuity of the inner product with respect to its arguments, Proposition 1.3.3, we have

$$\begin{aligned}(u, v) &= \lim_{n\to\infty} \left(\sum_{j=1}^{n} (u, v_j)\, v_j, \sum_{j=1}^{n} (v, v_j)\, v_j \right) \\ &= \lim_{n\to\infty} \sum_{j=1}^{n} (u, v_j)\, \overline{(v, v_j)} \\ &= \sum_{j=1}^{\infty} (u, v_j)\, \overline{(v, v_j)}.\end{aligned}$$

(b) $\Longrightarrow$ (c). Taking $u = v$ in (1.3.6), we obtain (1.3.7).
(c) $\Longrightarrow$ (d). For any $v \in V$, we have

$$\left\| v - \sum_{j=1}^{n} (v, v_j)\, v_j \right\|^2 = \|v\|^2 - \sum_{j=1}^{n} |(v, v_j)|^2 \to 0 \quad \text{as } n \to \infty,$$

by Parseval's equality (1.3.7). Hence, the subspace span $\{v_j\}_{j=1}^{\infty}$ is dense in V.

(d) $\Longrightarrow$ (e). If $(v, v_j) = 0$, $j \geq 1$, then v is orthogonal to the subspace span $\{v_j\}_{j=1}^{\infty}$. Since the inner product is continuous with respect to its arguments and since the subspace span $\{v_j\}_{j=1}^{\infty}$ is dense in V, we see that v is orthogonal to any element in V. In particular, $(v, v) = 0$, implying $v = 0$.

(e) $\Longrightarrow$ (a). For any $v \in V$, consider the element

$$w = v - \sum_{j=1}^{\infty} (v, v_j)\, v_j \in V.$$

For any $i \geq 1$, again by the continuity of the inner product, we have

$$(w, v_i) = (v, v_i) - \lim_{n\to\infty} \sum_{j=1}^{n} (v, v_j)\, \overline{(v_j, v_i)} = (v, v_i) - (v, v_i) = 0.$$

Hence, by statement (e), $w = 0$, i.e.,

$$v = \sum_{j=1}^{\infty} (v, v_j)\, v_j, \tag{1.3.8}$$

and statement (a) is valid. □

The advantage of using an orthogonal or an orthonormal basis is that it is easy to decompose a vector as a linear combination of the basis elements (see (1.3.8)). From the proof of the part "(e) $\Longrightarrow$ (a)" for Theorem 1.3.12 we see that an orthonormal basis $\{v_j\}_{j=1}^{\infty}$ is also a Schauder basis: for each $v \in V$, we have the formula (1.3.8).

We now introduce a result useful in the theory of Fourier series.

Theorem 1.3.13 *The functions*

$$\frac{1}{\sqrt{2\pi}},\ \frac{1}{\sqrt{\pi}} \cos jx,\ \frac{1}{\sqrt{\pi}} \sin jx,\ \ j = 1, 2, \cdots \tag{1.3.9}$$

form an orthonormal basis for the space $L^2(-\pi, \pi)$.

Proof. From Corollary 3.1.4 of Chapter 3, we note the following result about approximation by trigonometric polynomials: Suppose $f \in C([-\pi, \pi])$ satisfies $f(-\pi) = f(\pi)$. Then for any $\varepsilon > 0$, there exists a trigonometric polynomial q_n of a degree $n = n_\varepsilon$ such that

$$\max_{-\pi \leq x \leq \pi} |f(x) - q_n(x)| < \varepsilon.$$

Now consider any given $v \in L^2(-\pi, \pi)$ and $\varepsilon > 0$. By the density of $C([-\pi, \pi])$ in $L^2(-\pi, \pi)$ with respect to the $L^2(-\pi, \pi)$-norm, we have the existence of a function $v_{(1)} \in C([-\pi, \pi])$ such that

$$\|v - v_{(1)}\|_{L^2(-\pi,\pi)} < \varepsilon.$$

We then modify $v_{(1)}$ in neighborhoods of the end points $\pm\pi$ to construct $v_{(2)} \in C([-\pi,\pi])$ with the properties $v_{(2)}(-\pi) = v_{(2)}(\pi)$ and

$$\|v_{(1)} - v_{(2)}\|_{L^2(-\pi,\pi)} < \varepsilon.$$

For the function $v_{(2)}$, there exists a trigonometric polynomial q_n such that

$$\max_{-\pi\le x\le\pi} |v_{(2)}(x) - q_n(x)| < \frac{\varepsilon}{\sqrt{2\pi}}.$$

Then,

$$\|v_{(2)} - q_n\|_{L^2(-\pi,\pi)} < \varepsilon.$$

By the triangle inequality,

$$\begin{aligned}\|v - q_n\|_{L^2(-\pi,\pi)} &\le \|v - v_{(1)}\|_{L^2(-\pi,\pi)} + \|v_{(1)} - v_{(2)}\|_{L^2(-\pi,\pi)} \\ &\quad + \|v_{(1)} - q_n\|_{L^2(-\pi,\pi)} \\ &< 3\,\varepsilon.\end{aligned}$$

Hence, the subspace of the trigonometric polynomials is dense in $L^2(-\pi,\pi)$. The functions listed in (1.3.9) are clearly orthonormal. By Theorem 1.3.12, these functions form an orthonormal basis for the space $L^2(-\pi,\pi)$. □

Example 1.3.14 Let us show that the sequence

$$e_0(x) = \frac{1}{\sqrt{\pi}}, \quad e_k(x) = \sqrt{\frac{2}{\pi}}\cos(kx),\ k \ge 1$$

is an orthonormal basis in $L^2(0,\pi)$.

First, using the identity $(\cos x)^k = \left[(e^{ix} + e^{-ix})/2\right]^k$ we can show that any power $(\cos x)^k$ is a linear combination of the elements from $\{e_k(x)\}$. For any $f \in C[0,\pi]$, define $g \in C[-1,1]$ by $g(t) = f(\arccos t)$. Then apply Weierstrass' theorem (Theorem 3.1.1 from Chapter 3) to approximate $g(t)$ by a sequence of polynomials $\{p_n(t)\}$. The sequence $\{p_n(\cos x)\}$ approximates f. Finally we use the fact that $C[0,\pi]$ is dense in $L^2(0,\pi)$. □

Example 1.3.15 Let $V = L^2(-\pi,\pi)$. By Theorem 1.3.13, for any $v \in L^2(-\pi,\pi)$, we have the Fourier series expansion

$$v(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} \left[a_j \cos(jx) + b_j \sin(jx)\right],$$

where

$$\begin{aligned}a_j &= \frac{1}{\pi}\int_{-\pi}^{\pi} v(x)\,\cos(jx)\,dx, \quad j \ge 0, \\ b_j &= \frac{1}{\pi}\int_{-\pi}^{\pi} v(x)\,\sin(jx)\,dx, \quad j \ge 1.\end{aligned}$$

Also we have the ordinary Parseval's equality

$$\|v\|^2_{L^2(-\pi,\pi)} = \pi \left[\frac{|a_0|^2}{2} + \sum_{j=1}^{\infty} \left(|a_j|^2 + |b_j|^2 \right) \right], \tag{1.3.10}$$

followed from (1.3.7). □

When a non-orthogonal basis for an inner product space is given, there is a standard procedure to construct an orthonormal basis.

Theorem 1.3.16 *Let $\{w_n\}_{n\geq 1}$ be a set of linearly independent vectors in the inner product space V. Then there is an orthonormal system $\{v_n\}_{n\geq 1}$ with the property that*

$$\operatorname{span}\{w_n\}_{n=1}^{N} = \operatorname{span}\{v_n\}_{n=1}^{N} \quad \forall\, N \geq 1. \tag{1.3.11}$$

Proof. The proof is done inductively with the Gram-Schmidt method. For $N = 1$, define

$$v_1 = \frac{w_1}{\|w_1\|}$$

which satisfies $\|v_1\| = 1$. For $N \geq 2$, assume $\{v_n\}_{n=1}^{N-1}$ have been constructed with $(v_n, v_m) = \delta_{nm}$, $1 \leq n, m \leq N-1$, and

$$\operatorname{span}\{w_n\}_{n=1}^{N-1} = \operatorname{span}\{v_n\}_{n=1}^{N-1}.$$

Write

$$\tilde{v}_N = w_N + \sum_{n=1}^{N-1} \alpha_{N,n} v_n.$$

Now choose $\{\alpha_{N,n}\}_{n=1}^{N-1}$ by setting

$$(\tilde{v}_N, v_n) = 0, \quad 1 \leq n \leq N-1.$$

This implies

$$\alpha_{N,n} = -(w_N, v_n), \quad 1 \leq n \leq N-1.$$

This procedure "removes" from w_N the components in the directions of $v_1, \ldots, v_{N-1}$.

Finally, define

$$v_N = \frac{\tilde{v}_N}{\|\tilde{v}_N\|},$$

which is meaningful since $\tilde{v}_N \neq 0$. Then the sequence $\{v_n\}_{n=1}^{N}$ satisfies

$$(v_n, v_m) = \delta_{nm}, \quad 1 \leq n, m \leq N$$

and (1.3.11) holds. □

The Gram-Schmidt method can be used, e.g., to construct an orthonormal basis in $L^2(-1,1)$ for a polynomial space of certain degrees. As a result we obtain the well-known Legendre polynomials (after a proper scaling) which play an important role in some numerical analysis problems.

Example 1.3.17 Let us construct the first three orthonormal polynomials in $L^2(-1,1)$. For this purpose, we take

$$w_1(x) = 1, \quad w_2(x) = x, \quad w_3(x) = x^2.$$

Then easily,

$$v_1(x) = \frac{w_1(x)}{\|w_1\|} = \frac{1}{\sqrt{2}}.$$

To find $v_2(x)$, we write

$$\tilde{v}_2(x) = w_2(x) + \alpha_{2,1} v_1(x) = x + \frac{1}{\sqrt{2}}\alpha_{2,1}$$

and choose

$$\alpha_{2,1} = -(x, v_1(x)) = -\int_{-1}^{1} \frac{1}{\sqrt{2}}\, x\, dx = 0.$$

So $\tilde{v}_2(x) = x$, and

$$v_2(x) = \frac{\tilde{v}_2(x)}{\|\tilde{v}_2\|} = \sqrt{\frac{3}{2}}\, x.$$

Finally, we write

$$\tilde{v}_3(x) = w_3(x) + \alpha_{3,1} v_1(x) + \alpha_{3,2} v_2(x) = x^2 + \frac{1}{\sqrt{2}}\alpha_{3,1} + \sqrt{\frac{3}{2}}\alpha_{3,2} x.$$

Then

$$\alpha_{3,1} = -(w_3, v_1) = -\int_{-1}^{1} x^2 \frac{1}{\sqrt{2}}\, dx = -\frac{\sqrt{2}}{3},$$

$$\alpha_{3,2} = -(w_3, v_2) = -\int_{-1}^{1} x^2 \sqrt{\frac{3}{2}}\, x\, dx = 0.$$

Hence

$$\tilde{v}_3(x) = x^2 - \frac{1}{3}.$$

Since $\|\tilde{v}_3\|^2 = \frac{8}{45}$, we have

$$v_3(x) = \frac{3}{2}\sqrt{\frac{5}{2}}\left(x^2 - \frac{1}{3}\right).$$

The fourth orthonormal polynomial is

$$v_4(x) = \frac{1}{2}\sqrt{\frac{7}{2}}\left(5x^3 - 3x\right).$$

Graphs of these first four Legendre polynomials are given in Figure 1.2. □

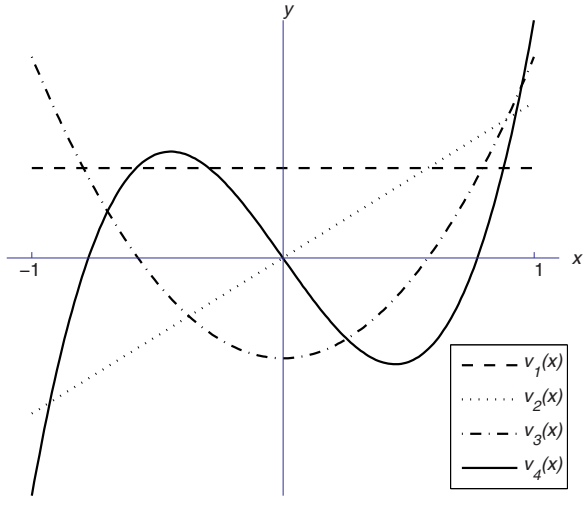

FIGURE 1.2. Graphs on $[-1, 1]$ of the orthonormal Legendre polynomials of degrees 0, 1, 2, 3

As is evident from Example 1.3.17, it is cumbersome to construct orthonormal (or orthogonal) polynomials by the Gram–Schmidt procedure. Fortunately, for many important cases of the weighted function $w(x)$ and integration interval (a, b), formulas of orthogonal polynomials in the weighted space $L^2_w(a, b)$ are known; see Section 3.5.

Exercise 1.3.1 Suppose $u \neq 0$ and $v \neq 0$. Derive the following equality

$$\|u\|\,\|v\| \pm (u, v) = \frac{1}{2}\,\|u\|\,\|v\| \left\| \frac{u}{\|u\|} \pm \frac{v}{\|v\|} \right\|^2$$

and use it to give another proof of Theorem 1.3.2 for the real case.

Exercise 1.3.2 Prove the Schwarz inequality for a complex inner product space V.
Hint: For $u, v \in V$, define an angle $\theta \in \mathbb{R}$ by the relation

$$(u, v) = |(u, v)|\, e^{i\,\theta}.$$

Then consider the function $\phi(t) = (u + t\, e^{i\,\theta} v, u + t\, e^{i\,\theta} v)$, which is quadratic and non-negative in t.

Exercise 1.3.3 Given an inner product, show that the formula $\|v\| = \sqrt{(v, v)}$ defines a norm.

Exercise 1.3.4 Show that in an inner product space, $\|u+v\| = \|u\| + \|v\|$ for some $u, v \in V$ if and only if u and v are non-negatively linearly dependent (i.e., for some $c_0 \geq 0$, either $u = c_0 v$ or $v = c_0 u$).

Exercise 1.3.5 In a real inner product space, show that

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2$$

implies $u \perp v$. For three non-zero vectors u, v and w satisfying

$$\|u+v+w\|^2 = \|u\|^2 + \|v\|^2 + \|w\|^2,$$

does there hold any orthogonality property?

Exercise 1.3.6 Assume $v_1, \cdots, v_n$ are mutually orthogonal non-zero vectors. Show that they are linearly independent.

Exercise 1.3.7 Derive the polarization identities (1.3.1) and (1.3.2) for a complex and real inner product space, respectively.

Exercise 1.3.8 Assume $f : \mathbb{R} \to \mathbb{R}$ is a continuous function, satisfying $f(\alpha) = f(\beta) + f(\alpha - \beta)$ for any $\alpha, \beta \in \mathbb{R}$, and $f(0) = 0$. Then $f(\alpha) = \alpha\, f(1)$.
Hint: Show first, $f(n\,\alpha) = n\, f(\alpha)$ for any integer n; then $f(1/2^n) = (1/2^n)\, f(1)$ for any integer $n \geq 0$; and finally, for any integer m, any non-negative integer n, $f(m\, 2^{-n}) = m\, f(2^{-n}) = (m\, 2^{-n})\, f(1)$. Represent any rational as a finite sum $q = \sum_i m_i\, 2^{-i}$, obtaining $f(q) = \sum_i f(m_i\, 2^{-i}) = \sum_i m_i\, 2^{-i}\, f(1) = q\, f(1)$. Use the density of the rational numbers in $\mathbb{R}$ and the continuity of f.

Exercise 1.3.9 The norms $\|\cdot\|_p$, $1 \leq p \leq \infty$, over the space $\mathbb{R}^d$ are defined in Example 1.2.3. Find all the values of p, for which the norm $\|\cdot\|_p$ is induced by an inner product.
Hint: Apply Theorem 1.3.4.

Exercise 1.3.10 Let $w_1, \ldots, w_d$ be positive constants. Show that the formula

$$(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{d} w_i x_i y_i$$

defines an inner product on $\mathbb{R}^d$. This is an example of a weighted inner product. What happens if we only assume $w_i \geq 0$, $1 \leq i \leq d$?

Exercise 1.3.11 Let $A \in \mathbb{R}^{d \times d}$ be a symmetric, positive definite matrix and let $(\cdot, \cdot)$ be the Euclidean inner product on $\mathbb{R}^d$. Show that the quantity $(A\boldsymbol{x}, \boldsymbol{y})$ defines an inner product on $\mathbb{R}^d$.

Exercise 1.3.12 Prove that the orthogonal complement of a subset is a closed subspace.

Exercise 1.3.13 Let V_0 be a subset of a Hilbert space V. Show that the following statements are equivalent:

(a) V_0 is dense in V, i.e., for any $v \in V$, there exists $\{v_n\}_{n\geq 1} \subset V_0$ such that $\|v - v_n\|_V \to 0$ as $n \to \infty$.

(b) $V_0^{\perp} = \{0\}$.

(c) If $u \in V$ satisfies $(u, v) = 0 \; \forall \, v \in V_0$, then $u = 0$.

(d) For every $0 \neq u \in V$, there is a $v \in V_0$ such that $(u, v) \neq 0$.

Exercise 1.3.14 On $C^1[a, b]$, define

$$(f, g)_* = f(a)g(a) + \int_a^b f'(x)g'(x)\, dx, \quad f, g \in C^1[a, b]$$

and $\|f\|_* = \sqrt{(f, f)_*}$. Show that

$$\|f\|_\infty \leq c\, \|f\|_* \quad \forall \, f \in C^1[a, b]$$

for a suitable constant c.

Exercise 1.3.15 Show that the sequence

$$e_k(x) = \sqrt{\frac{2}{\pi}}\, \sin(kx), \quad k \geq 1$$

is an orthonormal basis in $L^2(0, \pi)$.
Hint: Apply Theorem 1.3.12 and the result from Example 1.3.14 to the function $f(x) \sin x$.

Exercise 1.3.16 Extend the discussion of Example 1.3.15 to complex Fourier series. For any $v \in L^2(-\pi, \pi)$ with complex scalars, show the Fourier series expansion

$$v(x) = \sum_{n=-\infty}^{\infty} \alpha_n v_n(x) \quad \text{in } L^2(-\pi, \pi), \tag{1.3.12}$$

where

$$\alpha_n = (v, v_n) = \frac{1}{\sqrt{2\,\pi}} \int_{-\pi}^{\pi} v(x)\, e^{-inx} dx. \tag{1.3.13}$$

Give the Parseval equality.

Exercise 1.3.17 Continuing Exercise 1.3.16, consider the Fourier series (1.3.12) for a function $v \in C_p^m(2\pi)$ with $m \geq 2$. Show that

$$\left\| v - \sum_{n=-N}^{N} \alpha_n v_n \right\|_\infty \leq \frac{c_m(v)}{N^{m-1}}, \quad N \geq 1.$$

Hint: Use integration by parts in (1.3.13).

1.4 Spaces of continuously differentiable functions

Spaces of continuous functions and continuously differentiable functions were introduced in Example 1.1.2. In this section, we provide a more detailed review of these spaces.

From now on, unless stated explicitly otherwise, we will assume Ω to be a domain in $\mathbb{R}^d$, i.e., an open, bounded, connected subset of $\mathbb{R}^d$. A generic point in $\mathbb{R}^d$ is denoted by $\boldsymbol{x} = (x_1, \ldots, x_d)^T$. For multi-variable functions, it is convenient to use the multi-index notation for partial derivatives. A multi-index is an ordered collection of d non-negative integers, $\alpha = (\alpha_1, \ldots, \alpha_d)$. The quantity $|\alpha| = \sum_{i=1}^d \alpha_i$ is said to be the *length* of α. We will use the notation

$$\boldsymbol{x}^\alpha = \prod_{i=1}^d x_i^{\alpha_i}.$$

This is a monomial of degree $|\alpha|$. A general polynomial of degree less than or equal to n can be expressed by the formula

$$p(\boldsymbol{x}) = \sum_{|\alpha| \le n} a_\alpha \boldsymbol{x}^\alpha, \quad a_\alpha \in \mathbb{R},\ |\alpha| \le n.$$

For two multi-indices α and β, we define their sum $\alpha + \beta$ as the multi-index constructed by componentwise addition:

$$(\alpha + \beta)_i = \alpha_i + \beta_i, \quad 1 \le i \le d.$$

We write $\beta \le \alpha$ if $\beta_i \le \alpha_i$ for $i = 1, \cdots, d$. When $\beta \le \alpha$, we write $\alpha - \beta$ to mean the multi-index with the components $\alpha_i - \beta_i$, $1 \le i \le d$. Again for $\beta \le \alpha$, we define the binomial coefficient

$$\left(\begin{array}{c} \alpha \\ \beta \end{array} \right) = \prod_{i=1}^d \left(\begin{array}{c} \alpha_i \\ \beta_i \end{array} \right).$$

We recall that for two non-negative integers $m \le n$,

$$\left(\begin{array}{c} n \\ m \end{array} \right) = \frac{n!}{m!\,(n-m)!}. \tag{1.4.1}$$

By convention, $0! = 1$.

If v is an m-times differentiable function, then for any α with $|\alpha| \le m$,

$$\partial^\alpha v(\boldsymbol{x}) = \frac{\partial^{|\alpha|} v(\boldsymbol{x})}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

is the α^{th} order partial derivative. This is a handy notation for partial derivatives. Some examples are

$$\frac{\partial v}{\partial x_1} = \partial^\alpha v \quad \text{for } \alpha = (1, 0, \ldots, 0),$$
$$\frac{\partial^d v}{\partial x_1 \cdots \partial x_d} = \partial^\alpha v \quad \text{for } \alpha = (1, 1, \ldots, 1).$$

By convention, when $\alpha = (0, \ldots, 0)$, $\partial^\alpha v = v$. The set of all the derivatives of order m of a function v can be written as $\{\partial^\alpha v \mid |\alpha| = m\}$. For low order partial derivatives, there are other commonly used notations; e.g., the partial derivative $\partial v/\partial x_i$ is also written as $\partial_{x_i} v$, or $\partial_i v$, or $v_{,x_i}$, or $v_{,i}$.

The space $C(\Omega)$ consists of all real-valued functions which are continuous on Ω. Since Ω is open, a function from the space $C(\Omega)$ is not necessarily bounded. For example, with $d = 1$ and $\Omega = (0, 1)$, the function $v(x) = 1/x$ is continuous but unbounded on $(0, 1)$. Indeed, a function from the space $C(\Omega)$ can behave "nastily" as the variable approaches the boundary of Ω. Usually, it is more convenient to deal with continuous functions which are continuous up to the boundary. Let $C(\overline{\Omega})$ be the space of functions which are *uniformly continuous* on Ω. (It is important that Ω be a bounded set.) Any function in $C(\overline{\Omega})$ is bounded. The notation $C(\overline{\Omega})$ is consistent with the fact that a uniformly continuous function on Ω has a unique continuous extension to $\overline{\Omega}$. The space $C(\overline{\Omega})$ is a Banach space with its canonical norm

$$\|v\|_{C(\overline{\Omega})} = \sup\{|v(\boldsymbol{x})| \mid \boldsymbol{x} \in \Omega\} \equiv \max\{|v(\boldsymbol{x})| \mid \boldsymbol{x} \in \overline{\Omega}\}.$$

We have $C(\overline{\Omega}) \subset C(\Omega)$, and the inclusion is proper, i.e. there are functions $v \in C(\Omega)$ which cannot be extended to a continuous function on $\overline{\Omega}$. A simple example is $v(x) = 1/x$ on $(0, 1)$.

We use $\mathbb{Z}$ for the set of all the integers, and denote by $\mathbb{Z}_+$ the set of non-negative integers. For any $m \in \mathbb{Z}_+$, $C^m(\Omega)$ is the space of functions which, together with their derivatives of order less than or equal to m, are continuous on Ω; that is,

$$C^m(\Omega) = \{v \in C(\Omega) \mid \partial^\alpha v \in C(\Omega) \text{ for } |\alpha| \le m\}.$$

This is a linear space. The notation $C^m(\overline{\Omega})$ denotes the space of functions which, together with their derivatives of order less than or equal to m, are continuous up to the boundary:

$$C^m(\overline{\Omega}) = \{v \in C(\overline{\Omega}) \mid \partial^\alpha v \in C(\overline{\Omega}) \text{ for } |\alpha| \le m\}.$$

The space $C^m(\overline{\Omega})$ is a Banach space with the norm

$$\|v\|_{C^m(\overline{\Omega})} = \max_{|\alpha| \le m} \|\partial^\alpha v\|_{C(\overline{\Omega})}.$$

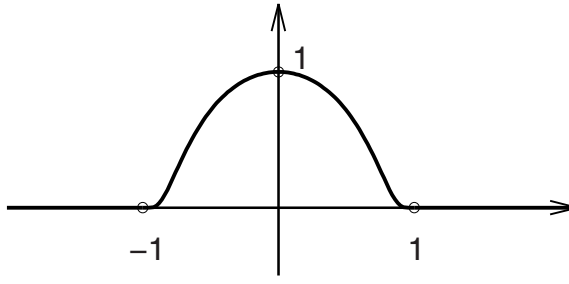

FIGURE 1.3. An infinitely smooth function with compact support

Algebraically, $C^m(\overline{\Omega}) \subset C^m(\Omega)$. When $m = 0$, we usually write $C(\Omega)$ and $C(\overline{\Omega})$ instead of $C^0(\Omega)$ and $C^0(\overline{\Omega})$. We set

$$C^\infty(\Omega) = \bigcap_{m=0}^{\infty} C^m(\Omega) \equiv \{v \in C(\Omega) \mid v \in C^m(\Omega) \ \ \forall\, m \in \mathbb{Z}_+\},$$

$$C^\infty(\overline{\Omega}) = \bigcap_{m=0}^{\infty} C^m(\overline{\Omega}) \equiv \{v \in C(\overline{\Omega}) \mid v \in C^m(\overline{\Omega}) \ \ \forall\, m \in \mathbb{Z}_+\}.$$

These are spaces of infinitely differentiable functions.

Given a function v on Ω, its support is defined to be

$$\operatorname{supp}(v) = \overline{\{\boldsymbol{x} \in \Omega \mid v(\boldsymbol{x}) \neq 0\}}.$$

We say that v has a *compact support* if $\operatorname{supp}(v)$ is a proper subset of Ω. Thus, if v has a compact support, then there is a neighboring open strip about the boundary $\partial\Omega$ such that v is zero on the part of the strip that lies inside Ω. Later on, we need the space

$$C_0^\infty(\Omega) = \{v \in C^\infty(\Omega) \mid \operatorname{supp}(v) \text{ is a proper subset of } \Omega\}.$$

Obviously, $C_0^\infty(\Omega) \subset C^\infty(\overline{\Omega})$. In the case Ω is an interval such that $\Omega \supset [-1, 1]$, a standard example of a $C_0^\infty(\Omega)$ function is

$$u_0(x) = \begin{cases} e^{x^2/(x^2-1)}, & |x| < 1, \\ 0, & \text{otherwise.} \end{cases}$$

A graph of the function $u_0(x)$ is shown in Figure 1.3.

1.4.1 Hölder spaces

A function v defined on Ω is said to be *Lipschitz continuous* if for some constant c, there holds the inequality

$$|v(\boldsymbol{x}) - v(\boldsymbol{y})| \leq c\,\|\boldsymbol{x} - \boldsymbol{y}\| \qquad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$

In this formula, $\|\boldsymbol{x}-\boldsymbol{y}\|$ denotes the standard Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. The smallest possible constant in the above inequality is called the *Lipschitz constant* of v, and is denoted by $\mathrm{Lip}(v)$. The Lipschitz constant is characterized by the relation

$$\mathrm{Lip}(v) = \sup\left\{ \frac{|v(\boldsymbol{x})-v(\boldsymbol{y})|}{\|\boldsymbol{x}-\boldsymbol{y}\|} \;\Big|\; \boldsymbol{x},\boldsymbol{y}\in\Omega,\ \boldsymbol{x}\neq\boldsymbol{y} \right\}.$$

More generally, a function v is said to be *Hölder continuous* with exponent $\beta\in(0,1]$ if for some constant c,

$$|v(\boldsymbol{x})-v(\boldsymbol{y})| \le c\,\|\boldsymbol{x}-\boldsymbol{y}\|^{\beta} \quad \text{for } \boldsymbol{x},\boldsymbol{y}\in\Omega.$$

The Hölder space $C^{0,\beta}(\overline{\Omega})$ is defined to be the subspace of $C(\overline{\Omega})$ functions that are Hölder continuous with the exponent β. With the norm

$$\|v\|_{C^{0,\beta}(\overline{\Omega})} = \|v\|_{C(\overline{\Omega})} + \sup\left\{ \frac{|v(\boldsymbol{x})-v(\boldsymbol{y})|}{\|\boldsymbol{x}-\boldsymbol{y}\|^{\beta}} \;\Big|\; \boldsymbol{x},\boldsymbol{y}\in\Omega,\ \boldsymbol{x}\neq\boldsymbol{y} \right\}$$

the space $C^{0,\beta}(\overline{\Omega})$ becomes a Banach space. When $\beta=1$, the Hölder space $C^{0,1}(\overline{\Omega})$ consists of all the Lipschitz continuous functions.

For $m\in\mathbb{Z}_+$ and $\beta\in(0,1]$, we similarly define the Hölder space

$$C^{m,\beta}(\overline{\Omega}) = \left\{ v\in C^m(\overline{\Omega}) \mid \partial^\alpha v\in C^{0,\beta}(\overline{\Omega})\ \forall\,\alpha \text{ with } |\alpha|=m \right\};$$

this is a Banach space with the norm

$$\|v\|_{C^{m,\beta}(\overline{\Omega})} = \|v\|_{C^m(\overline{\Omega})} + \sum_{|\alpha|=m} \sup\left\{ \frac{|\partial^\alpha v(\boldsymbol{x})-\partial^\alpha v(\boldsymbol{y})|}{\|\boldsymbol{x}-\boldsymbol{y}\|^{\beta}} \;\Big|\; \boldsymbol{x},\boldsymbol{y}\in\Omega,\ \boldsymbol{x}\neq\boldsymbol{y} \right\}.$$

For $m\in\mathbb{Z}_+$ and $0<\alpha<\beta\le 1$, we have the obvious relations

$$C^{m,1}(\overline{\Omega}) \subset C^{m,\beta}(\overline{\Omega}) \subset C^{m,\alpha}(\overline{\Omega}) \subset C^m(\overline{\Omega}),$$

valid for any open set $\Omega\subset\mathbb{R}^d$. When Ω is bounded and convex, we also have

$$C^{m+1}(\overline{\Omega}) \subset C^{m,1}(\overline{\Omega}).$$

Exercise 1.4.1 Show that $C(\overline{\Omega})$ with the norm $\|v\|_{C(\overline{\Omega})}$ is a Banach space.

Exercise 1.4.2 Show that the space $C^1(\overline{\Omega})$ with the norm $\|v\|_{C(\overline{\Omega})}$ is not a Banach space.

Exercise 1.4.3 Let $\Omega\subset\mathbb{R}^d$ be a non-empty open set. Construct a function that belongs to the space $C_0^\infty(\Omega)$.

Exercise 1.4.4 Let $v_n(x) = (\sin nx)/n$. Show that $v_n \to 0$ in $C^{0,\beta}[0,1]$ for any $\beta \in (0,1)$, but $v_n \not\to 0$ in $C^{0,1}[0,1]$.

Exercise 1.4.5 Discuss whether it is meaningful to use the Hölder space $C^{0,\beta}(\overline{\Omega})$ with $\beta > 1$.

Exercise 1.4.6 Consider $v(s) = s^\alpha$ for some $0 < \alpha < 1$. Show that $v \in C^{0,\beta}[0,1]$ for any $\beta \in (0,\alpha]$; moreover, $\|v\|_{C^{0,\alpha}[0,1]} = 2$.
Hint: To compute the maximum value of $|s^\alpha - t^\alpha|/|s-t|^\alpha$ for $s,t \in [0,1]$ and $s \neq t$, it is sufficient to consider the case $s > t$. For fixed $z > 1$, define $f(\alpha) = (z^\alpha - 1)/(z-1)^\alpha$. Show that $f(\alpha)$ is an increasing function of $\alpha \in (0,1]$, and so $f(\alpha) \leq f(1) = 1$. On the other hand, with $t = 0$, $|s^\alpha - 0|/|s-0|^\alpha = 1$.

Exercise 1.4.7 Let $m \geq 0$ be an integer, $\beta \in (0,1]$. Show that $C^{m+1}(\overline{\Omega}) \subset C^{m,\beta}(\overline{\Omega})$, and moreover, "$v_n \to v$ in $C^{m+1}(\overline{\Omega})$" implies "$v_n \to v$ in $C^{m,\beta}(\overline{\Omega})$".

Exercise 1.4.8 This and the next two exercises let the reader get familiar with some formulas and results involving the multi-index notation. Show the generalization of the formula (1.4.1) to binomial numbers of two multi-indices:

$$\left(\begin{array}{c} \alpha \\ \beta \end{array} \right) = \frac{\alpha!}{\beta!\,(\alpha-\beta)!}, \quad \beta \leq \alpha.$$

Exercise 1.4.9 Recall the binomial theorem:

$$(a+b)^n = \sum_{m=0}^{n} \left(\begin{array}{c} n \\ m \end{array} \right) a^m b^{n-m}.$$

Use this result to prove the multi-variate binomial formula

$$(\boldsymbol{x}+\boldsymbol{y})^\alpha = \sum_{\beta \leq \alpha} \left(\begin{array}{c} \alpha \\ \beta \end{array} \right) \boldsymbol{x}^\beta \boldsymbol{y}^{\alpha-\beta}, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d.$$

Exercise 1.4.10 We use the symbol $\mathbb{P}_n = \mathbb{P}_n(\mathbb{R}^d)$ or $\mathbb{P}_n(\Omega)$, $\Omega \subset \mathbb{R}^d$ a nonempty open set, for the space of polynomials of degree less than or equal to n in $\mathbb{R}^d$ or Ω. The dimension of the space $\mathbb{P}_n$ is the number of monomials $\boldsymbol{x}^\alpha$, $|\alpha| \leq n$. Show by an inductive argument that

$$\dim \mathbb{P}_n = \left(\begin{array}{c} n+d \\ d \end{array} \right).$$

Exercise 1.4.11 Consider the function

$$f(x) = \begin{cases} e^{-1/|x|}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

Prove

$$\lim_{x \to 0} f^{(m)}(x) = 0$$

for all integers $m \geq 0$.

Exercise 1.4.12 Let $0 < a < b < \infty$. Define

$$f(\boldsymbol{x}) = \begin{cases} \exp\bigl(1/(a^2 - \|\boldsymbol{x}\|^2) + 1/(\|\boldsymbol{x}\|^2 - b^2)\bigr), & a < \|\boldsymbol{x}\| < b, \\ 0, & \|\boldsymbol{x}\| \le a \text{ or } \|\boldsymbol{x}\| \ge b. \end{cases}$$

Show that $f \in C_0^\infty(\mathbb{R}^d)$.

Exercise 1.4.13 Construct a function $f \in C^\infty(\mathbb{R})$ such that $f(x) = 0$ for $x \le -1$, and $f(x) = 1$ for $x \ge 1$.

1.5 L^p spaces

Let $\Omega \subset \mathbb{R}^d$ be a non-empty open set. In the study of $L^p(\Omega)$ spaces, we identify functions (i.e. such functions are considered identical) which are equal a.e. on Ω. For $p \in [1, \infty)$, $L^p(\Omega)$ is the linear space of measurable functions $v : \Omega \to \mathbb{R}$ such that

$$\|v\|_{L^p(\Omega)} = \left[\int_\Omega |v(\boldsymbol{x})|^p dx\right]^{1/p} < \infty. \tag{1.5.1}$$

The space $L^\infty(\Omega)$ consists of all essentially bounded measurable functions $v : \Omega \to \mathbb{R}$,

$$\|v\|_{L^\infty(\Omega)} = \inf_{\text{meas}\,(\Omega')=0} \sup_{\boldsymbol{x} \in \Omega \setminus \Omega'} |v(\boldsymbol{x})| < \infty. \tag{1.5.2}$$

For $p = 1, 2, \infty$, it is quite straightforward to show $\|\cdot\|_{L^p(\Omega)}$ is a norm. For other values of p, in proving $\|\cdot\|_{L^p(\Omega)}$ is a norm, the main difficulty is to show the triangle inequality, known as the Minkowski inequality (Lemma 1.5.4 below). To prove the Minkowski inequality, we start with Young's inequality and then the Hölder inequality.

For $p \in [1, \infty]$, we define its conjugate q by the relation

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{1.5.3}$$

Here we adopt the convention $1/\infty = 0$. It is easy to see that $1 \le q \le \infty$. Moreover, $1 < q < \infty$ if $1 < p < \infty$, $q = 1$ if $p = \infty$, and $q = \infty$ if $p = 1$. For $p \ne 1, \infty$, we obtain from the defining relation (1.5.3) that its conjugate is given by the formula

$$q = \frac{p}{p-1}. \tag{1.5.4}$$

We first introduce Young's inequality.

Lemma 1.5.1 (Young's Inequality) *Let $a, b \ge 0$, $1 < p < \infty$, and q the conjugate of p. Then*

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}.$$

Proof. For any fixed $b \geq 0$, define a function

$$f(a) = \frac{a^p}{p} + \frac{b^q}{q} - ab$$

on $[0, \infty)$. From $f'(a) = 0$ we obtain $a = b^{1/(p-1)}$. We have $f(b^{1/(p-1)}) = 0$. Since $f(0) \geq 0$, $\lim_{a\to\infty} f(a) = \infty$ and f is continuous on $[0, \infty)$, we see that

$$\inf_{0\leq a<\infty} f(a) = f(b^{1/(p-1)}) = 0.$$

Hence Young's inequality holds. □

Some other ways of proving Young's inequality are given in Exercises 1.5.2–1.5.4.

It is usually useful to include a parameter in the Young inequality.

Lemma 1.5.2 (Modified Young's inequality) *Let $a, b \geq 0$, $\varepsilon > 0$, $1 < p < \infty$, and q the conjugate of p. Then*

$$ab \leq \frac{\varepsilon\, a^p}{p} + \frac{\varepsilon^{1-q} b^q}{q}.$$

Lemma 1.5.3 (Hölder's inequality) *Let $p \in [1, \infty]$ and q be its conjugate. Then for any $u \in L^p(\Omega)$ and any $v \in L^q(\Omega)$,*

$$\int_\Omega |u(\boldsymbol{x})\, v(\boldsymbol{x})|\, dx \leq \|u\|_{L^p(\Omega)} \|v\|_{L^q(\Omega)}.$$

Proof. The inequality is obviously true if $p = 1$ or ∞, or $\|u\|_{L^p(\Omega)} = 0$. For $p \in (1, \infty)$ and $\|u\|_{L^p(\Omega)} \neq 0$, we use the modified Young's inequality to obtain

$$\int_\Omega |u(\boldsymbol{x})\, v(\boldsymbol{x})|\, dx \leq \frac{\varepsilon}{p} \|u\|^p_{L^p(\Omega)} + \frac{\varepsilon^{1-q}}{q} \|v\|^q_{L^q(\Omega)} \quad \forall\, \varepsilon > 0.$$

Then we set $\varepsilon = \|v\|_{L^q(\Omega)} / \|u\|^{p-1}_{L^p(\Omega)}$; this choice of ε minimizes the value of the right-hand side of the above inequality. □

We are now ready to show the *Minkowski inequality*, i.e. the triangle inequality for $\|\cdot\|_{L^p(\Omega)}$ defined in (1.5.1)–(1.5.2).

Lemma 1.5.4 (Minkowski inequality)

$$\|u + v\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)} \quad \forall\, u, v \in L^p(\Omega),\ p \in [1, \infty].$$

Proof. The inequality is obviously true for $p = 1$ and ∞. Suppose $p \in (1, \infty)$. Applying the Hölder inequality, we have

$$\begin{aligned}
&\int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^p dx \\
&\qquad \le \int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^{p-1} |u(\boldsymbol{x})|\, dx + \int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^{p-1} |v(\boldsymbol{x})|\, dx \\
&\qquad \le \left[\int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^{(p-1)q} dx\right]^{1/q} \left[\|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}\right] \\
&\qquad = \left[\int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^p dx\right]^{1-1/p} \left[\|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}\right].
\end{aligned}$$

Therefore, the Minkowski inequality holds. □

The reference [118] provides a comprehensive study of inequalities in analysis.

Some basic properties of the L^p spaces are summarized in the following theorem.

Theorem 1.5.5 *Let Ω be an open bounded set in $\mathbb{R}^d$.*

(a) *For $p \in [1, \infty]$, $L^p(\Omega)$ is a Banach space.*

(b) *For $p \in [1, \infty]$, every Cauchy sequence in $L^p(\Omega)$ has a subsequence which converges pointwise a.e. on Ω.*

(c) *If $1 \le p \le q \le \infty$, then $L^q(\Omega) \subset L^p(\Omega)$,*

$$\|v\|_{L^p(\Omega)} \le \text{meas}\,(\Omega)^{1/p-1/q} \|v\|_{L^q(\Omega)} \quad \forall\, v \in L^q(\Omega),$$

and

$$\|v\|_{L^\infty(\Omega)} = \lim_{p\to\infty} \|v\|_{L^p(\Omega)} \quad \forall\, v \in L^\infty(\Omega).$$

(d) *If $1 \le p \le r \le q \le \infty$ and we choose $\theta \in [0, 1]$ such that*

$$\frac{1}{r} = \frac{\theta}{p} + \frac{(1-\theta)}{q},$$

then

$$\|v\|_{L^r(\Omega)} \le \|v\|^\theta_{L^p(\Omega)} \|v\|^{1-\theta}_{L^q(\Omega)} \quad \forall\, v \in L^q(\Omega).$$

In (c), when $q = \infty$, $1/q$ is understood to be 0. The result (d) is called an *interpolation property* of the L^p spaces. We can use the Hölder inequality to prove (c) and (d) (Exercise 1.5.5).

For $p \in (1, \infty)$, we have the following Clarkson inequalities. Let $u, v \in L^p(\Omega)$. If $2 \le p < \infty$, then

$$\left\|\frac{u+v}{2}\right\|^p_{L^p(\Omega)} + \left\|\frac{u-v}{2}\right\|^p_{L^p(\Omega)} \le \frac{1}{2}\|u\|^p_{L^p(\Omega)} + \frac{1}{2}\|v\|^p_{L^p(\Omega)}. \tag{1.5.5}$$

If $1 < p \le 2$, then

$$\left\| \frac{u+v}{2} \right\|_{L^p(\Omega)}^q + \left\| \frac{u-v}{2} \right\|_{L^p(\Omega)}^q \le \left[\frac{1}{2} \|u\|_{L^p(\Omega)}^p + \frac{1}{2} \|v\|_{L^p(\Omega)}^p \right]^{q-1}, \quad (1.5.6)$$

where $q = p/(p-1)$ is the conjugate exponent of p. A proof of these inequalities can be found in [1, Chapter 2].

Smooth functions are dense in $L^p(\Omega)$, $1 \le p < \infty$.

Theorem 1.5.6 *Let $\Omega \subset \mathbb{R}^d$ be an open set, $1 \le p < \infty$. Then the space $C_0^\infty(\Omega)$ is dense in $L^p(\Omega)$; in other words, for any $v \in L^p(\Omega)$, there exists a sequence $\{v_n\} \subset C_0^\infty(\Omega)$ such that*

$$\|v_n - v\|_{L^p(\Omega)} \to 0 \quad \text{as } n \to \infty.$$

For any $m \in \mathbb{Z}_+$, by noting the inclusions $C_0^\infty(\Omega) \subset C^m(\overline{\Omega}) \subset L^p(\Omega)$, we see that the space $C^m(\overline{\Omega})$ is also dense in $L^p(\Omega)$.

Exercise 1.5.1 Prove the modified Young's inequality by applying Young's inequality.

Exercise 1.5.2 Lemma 1.5.1 is a special case of the following general Young's inequality: Let $f : [0, \infty) \to [0, \infty)$ be a continuous, strictly increasing function such that $f(0) = 0$, $\lim_{x\to\infty} f(x) = \infty$. Denote by g the inverse function of f. For $0 \le x < \infty$, define

$$F(x) = \int_0^x f(t)\, dt, \qquad G(x) = \int_0^x g(t)\, dt.$$

Then

$$ab \le F(a) + F(b) \quad \forall\, a, b \ge 0,$$

and the equality holds if and only if $b = f(a)$. Prove this result and deduce Lemma 1.5.1 from it.

Exercise 1.5.3 Show that Young's inequality can be written equivalently as

$$a^\theta b^{1-\theta} \le \theta a + (1-\theta) b \quad \forall\, a, b \ge 0,\ \theta \in (0, 1).$$

Prove this inequality by using the convexity of the natural exponential function. Further, prove Hölder's inequality for $p \in (1, \infty)$ from this inequality by taking

$$\theta = \frac{1}{p}, \quad a = \frac{|u(\boldsymbol{x})|^p}{\|u\|_{L^p(\Omega)}^p}, \quad b = \frac{|v(\boldsymbol{x})|^q}{\|v\|_{L^q(\Omega)}^q}.$$

Exercise 1.5.4 Another method to prove the Young inequality is based on the fact that $\log x$ is a concave function so that

$$\log\left(\frac{a^p}{p} + \frac{b^q}{q} \right) \ge \frac{1}{p} \log a^p + \frac{1}{q} \log b^q.$$

Use this inequality to prove the Young inequality.

Exercise 1.5.5 Use the Hölder inequality to prove (c) and (d) of Theorem 1.5.5.

Exercise 1.5.6 Show the generalized Hölder inequality

$$\left| \int_\Omega v_1 \cdots v_m \, dx \right| \le \|v_1\|_{L^{p_1}(\Omega)} \cdots \|v_m\|_{L^{p_m}(\Omega)} \quad \forall\, v_i \in L^{p_i}(\Omega),\ 1 \le i \le m,$$

where the exponents $p_i > 0$ satisfy the relation $\sum_{i=1}^m 1/p_i = 1$.

Exercise 1.5.7 Prove the discrete analogue of Lemma 1.5.3, the Hölder inequality on $\mathbb{R}^d$:

$$|\boldsymbol{x} \cdot \boldsymbol{y}| \le \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q, \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,$$

where $1 \le p, q \le \infty$, and q is the conjugate of p. Then prove the Minkowski inequality on $\mathbb{R}^d$:

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p \le \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d,\ 1 \le p \le \infty.$$

Show that for $p \in (1, \infty)$, $\|\cdot\|_p$ defined by (1.2.2) is a norm on $\mathbb{R}^d$. The cases where $p = 1, 2, \infty$ have been considered in Exercise 1.2.6.

Exercise 1.5.8 Let $1 \le p, q < \infty$. Show that

$$\max_{a \le x \le b} |f(x)| \le (b-a)^{-1/p} \|f\|_{L^p(a,b)} + (b-a)^{(q-1)/q} \|f'\|_{L^q(a,b)} \quad \forall\, f \in C^1[a,b].$$

Hint: Start with $f(x) = f(t) + \int_t^x f'(s)\, ds$.

Exercise 1.5.9 The evolution of two functions f and g on $\mathbb{R}^d$ is defined by the formula

$$(f * g)(\boldsymbol{x}) = \int_{\mathbb{R}^d} f(\boldsymbol{y})\, g(\boldsymbol{x} - \boldsymbol{y})\, dy$$

whenever the integral exists a.e. Let $1 \le p \le \infty$. Show that if $f \in L^p(\mathbb{R}^d)$ and $g \in L^1(\mathbb{R}^d)$, then $f * g \in L^p(\mathbb{R}^d)$ and

$$\|f * g\|_{L^p(\mathbb{R}^d)} \le \|f\|_{L^p(\mathbb{R}^d)} \|g\|_{L^1(\mathbb{R}^d)}.$$

Hint: For $1 < p < \infty$, write

$$|(f * g)(\boldsymbol{x})| \le \int_{\mathbb{R}^d} \left[|f(\boldsymbol{y})|\, |g(\boldsymbol{x} - \boldsymbol{y})|^{1/p} \right] |g(\boldsymbol{x} - \boldsymbol{y})|^{1/q} dy$$

where q is the conjugate exponent of p, and apply Lemma 1.5.3.

Exercise 1.5.10 For $p \in [1, \infty]$, recall the definition (1.2.4) of the space ℓ^p. Prove the discrete Hölder inequality: Let $p \in (1, \infty)$, $1/p + 1/q = 1$, $u = (u_n)_{n \ge 1} \in \ell^p$ and $v = (v_n)_{n \ge 1} \in \ell^q$. Then

$$\sum_{n=1}^\infty u_n v_n \le \|u\|_{\ell^p} \|v\|_{\ell^q}.$$

Hint: Mimic the proof of Lemma 1.5.3, or apply Lemma 1.5.3 through choosing piecewise constant functions $f(x)$ and $g(x)$, $0 \le x < \infty$, with values from the components of u and v.

Exercise 1.5.11 Continuing the previous exercise, prove the discrete Minkowski inequality:

$$\|u+v\|_{\ell^p} \le \|u\|_{\ell^p} + \|v\|_{\ell^p} \quad \forall\, u, v \in \ell^p,\ 1 \le p \le \infty.$$

Exercise 1.5.12 Show by examples that when Ω is unbounded, for $1 \le p < q \le \infty$, we have neither $L^p(\Omega) \subset L^q(\Omega)$ nor $L^q(\Omega) \subset L^p(\Omega)$.

1.6 Compact sets

There are several definitions of the concept of compact set, most being equivalent in the setting of a normed space.

Definition 1.6.1 (a) *Let S be a subset of a normed space V. We say S has an* open covering *by a collection of open sets $\{U_\alpha \mid \alpha \in \Lambda\}$, Λ an index set, if*

$$S \subset \bigcup_{\alpha\in\Lambda} U_\alpha\,.$$

We say S is compact *if for every open covering $\{U_\alpha\}$ of S, there is a finite subcover $\{U_{\alpha_j} \mid j = 1, \ldots, m\} \subset \{U_\alpha \mid \alpha \in \Lambda\}$ which also covers S.*

(b) *Equivalently, S is* compact *if every sequence $\{v_j\} \subset S$ contains a convergent subsequence $\{v_{j_k}\}$ which converges to an element $v \in S$.*

(c) *If S is a set for which the closure $\overline{S}$ is compact, we say S is* precompact.

Part (a) of the definition is the general definition of compact set, valid in general topological spaces; and (b) is the usual form used in metric spaces (spaces with a distance function defining the topology of the space, see Exercise 1.2.16). In a general topological space, part (b) defines sequential compactness. In every normed space, a compact set is both closed and bounded.

For finite dimensional spaces, the compact sets are readily identified.

Theorem 1.6.2 (Heine-Borel Theorem) *Let V be a finite dimensional normed space, and let S be a subset of V. Then S is compact if and only if S is both closed and bounded.*

A proof of this result can be found in most textbooks on advanced calculus; for example, see [7, Theorems 3–38, 3–40]. Indeed, a normed space is finite dimensional if and only if any closed and bounded set in the space is compact. For a proof of this result, see e.g. [49, Section 1.4].

For infinite dimensional normed spaces, it is more difficult to identify the compact sets. The results are dependent on the properties of the norm being used. We give an important result for the space of continuous functions $C(D)$ with the uniform norm $\|\cdot\|_\infty$, with some set $D \subset \mathbb{R}^d$. A proof is given in [135, p. 27].

Theorem 1.6.3 (Arzela-Ascoli Theorem) *Let $S \subset C(D)$, with $D \subset \mathbb{R}^d$ closed and bounded. Suppose that the functions in S are uniformly bounded and equicontinuous over D, meaning that*

$$\sup_{f \in S} \|f\|_\infty < \infty$$

and

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq c_S(\varepsilon) \text{ for } \|\boldsymbol{x} - \boldsymbol{y}\| \leq \varepsilon \quad \forall\, f \in S,\ \forall\, \boldsymbol{x}, \boldsymbol{y} \in D,$$

with $c_S(\varepsilon) \to 0$ as $\varepsilon \to 0$. Then S is precompact in $C(D)$.

In Chapter 7, we review compact embedding results for Sobolev spaces, and these provide examples of compact sets in Sobolev spaces.

Exercise 1.6.1 Show that in a normed space, a compact set is closed and bounded.

Exercise 1.6.2 Show that the intersection of a family of compact sets and the union of a finite number of compact sets are compact. Construct an example to show that the union of an infinite number of compact sets may fail to be compact.

Exercise 1.6.3 Use Heine-Borel Theorem to show that any finite-dimensional normed space is complete.

Exercise 1.6.4 In a normed space, any closed subset of a compact set is compact.

Exercise 1.6.5 Use the Arzala-Ascoli theorem to show that a bounded set in $C^1[a, b]$ is precompact in $C[a, b]$.

Suggestion for Further Reading.

Detailed discussions of normed spaces, Banach spaces, Hilbert spaces, linear operators on normed spaces and their properties are found in most textbooks on Functional Analysis; see for example Conway [58], Hutson and Pym [128], Kantorovich and Akilov [135], Zeidler [249], [250].

In this work, we emphasize the ability to understand and correctly apply results from functional analysis in order to analyze various numerical methods and procedures. An important pioneering paper which advocated and developed this approach to numerical analysis is that of L.V. Kantorovich [134], appearing in 1948; and much of the work of this paper appears in expanded form in the book of Kantorovich and Akilov [135, Chaps 14–18]. Another associated and influential work is the text of Kantorovich and Krylov [136] which appeared in several editions over a 30 year period. Other important general texts which set the study of numerical analysis within a framework of functional analysis include Collatz [56], Cryer [60], Lebedev [154], and Linz [157].

2
Linear Operators on Normed Spaces

Many of the basic problems of applied mathematics share the property of *linearity*, and linear spaces and linear operators provide a general and useful framework for the analysis of such problems. More complicated applications often involve nonlinear operators, and a study of linear operators also offers some useful tools for the analysis of nonlinear operators. In this chapter we review some basic results on linear operators, and we give some illustrative applications to obtain results in numerical analysis. Some of the results are quoted without proof; and usually the reader can find detailed proofs of the results in a standard textbook on functional analysis, e.g. see Conway [58], Kantorovich and Akilov [135], and Zeidler [249], [250].

Linear operators are used in expressing mathematical problems, often leading to equations to be solved or to functions to be optimized. To examine the theoretical solvability of a mathematical problem and to develop numerical methods for its solution, we must know additional properties about the operators involved in our problem. The most important such properties in applied mathematics involve one of the following concepts or some mix of them.

- Closeness to a problem whose solvability theory is known. The *Geometric Series Theorem* given in Section 2.3 is the basis of most results for linear operator equations in this category.

- Closeness to a finite dimensional problem. One variant of this leads to the theory of *completely continuous* or *compact* linear operators, which is taken up in Section 2.8.

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_2,
© Springer Science + Business Media, LLC 2009

- Arguments based on finding the minimum of a function, with the point at which the minimum is attained being the solution to the problem under study. The function being minimized is sometimes called an objective function in optimization theory or an energy function in mechanics applications. This is taken up in later chapters, but some of its framework is provided in the material of this chapter.

There are other important means of examining the solvability of mathematical problems in applied mathematics, based on Fourier analysis, complex analysis, positivity of an operator within the context of partially order linear spaces, and other techniques. However, we make only minimal use of such tools in this text.

2.1 Operators

Given two sets V and W, an *operator* T from V to W is a rule which assigns to each element in a subset of V a unique element in W. The *domain* $\mathcal{D}(T)$ of T is the subset of V where T is defined:

$$\mathcal{D}(T) = \{v \in V \mid T(v) \text{ is defined}\},$$

and the *range* $\mathcal{R}(T)$ of T is the set of the elements in W generated by T:

$$\mathcal{R}(T) = \{w \in W \mid w = T(v) \text{ for some } v \in \mathcal{D}(T)\}.$$

It is also useful to define the *null set*, the set of the zeros of the operator:

$$\mathcal{N}(T) = \{v \in V \mid T(v) = 0\}.$$

An operator is sometimes also called a mapping, a transformation, or a function. Usually the domain $\mathcal{D}(T)$ is understood to be the whole set V, unless it is stated explicitly otherwise. Also, from now on, we will assume both V and W are linear spaces, as this suffices in the rest of the book.

Addition and scalar multiplication of operators are defined similarly to that of ordinary functions. Let S and T be operators mapping from V to W. Then $S+T$ is an operator from V to W with the domain $\mathcal{D}(S) \cap \mathcal{D}(T)$ and the rule

$$(S+T)(v) = S(v) + T(v) \quad \forall\, v \in \mathcal{D}(S) \cap \mathcal{D}(T).$$

Let $\alpha \in \mathbb{K}$. Then αT is an operator from V to W with the domain $\mathcal{D}(T)$ and the rule

$$(\alpha T)(v) = \alpha\, T(v) \quad \forall\, v \in \mathcal{D}(T).$$

Definition 2.1.1 *An operator $T : V \to W$ is said to be* one-to-one *or* injective *if*

$$v_1 \neq v_2 \quad \Longrightarrow \quad T(v_1) \neq T(v_2). \tag{2.1.1}$$

The operator is said to map V onto *W or is called* surjective *if $\mathcal{R}(T) = W$. If T is both injective and surjective, it is called a* bijection *from V to W.*

Evidently, when $T : V \to W$ is bijective, we can define its inverse $T^{-1} : W \to V$ by the rule

$$v = T^{-1}(w) \quad \Longleftrightarrow \quad w = T(v).$$

More generally, if $T : V \to W$ is one-to-one, we can define its inverse from $\mathcal{R}(T) \subset W$ to V by using the above rule.

Example 2.1.2 Let V be a linear space. The *identity operator* $I : V \to V$ is defined by

$$I(v) = v \quad \forall\, v \in V.$$

It is a bijection from V to V; and moreover, its inverse is also the identity operator. □

Example 2.1.3 Let $V = \mathbb{R}^n$, $W = \mathbb{R}^m$, and $L(\boldsymbol{v}) = A\boldsymbol{v}$, $\boldsymbol{v} \in \mathbb{R}^n$, where $A = (a_{ij}) \in \mathbb{R}^{m\times n}$ is a real matrix and $A\boldsymbol{v}$ denotes matrix-vector multiplication. From results in linear algebra, the operator L is injective if and only if $\operatorname{rank}(A) = n$; and L is surjective if and only if $\operatorname{rank}(A) = m$. Recall that the rank of a matrix is the maximal number of independent column vectors, that is also the maximal number of independent row vectors.

The same conclusion holds for complex spaces $V = \mathbb{C}^n$, $W = \mathbb{C}^m$, and complex matrix $A \in \mathbb{C}^{m\times n}$. □

Example 2.1.4 We consider the differentiation operator d/dx from $V = C[0,1]$ to $W = C[0,1]$ defined by

$$\frac{d}{dx} : v \mapsto v' \quad \text{for } v \in C^1[0,1].$$

We take the domain of the operator, $\mathcal{D}(d/dx)$, to be $C^1[0,1]$ which is a proper subspace of $C[0,1]$. It can be verified that the differentiation operator is a surjection, $\mathcal{R}(d/dx) = C[0,1]$. The differentiation operator is not injective, and its null set is the set of constant functions. □

Example 2.1.5 Although the differentiation operator d/dx is not injective from $C^1[0,1]$ to $C[0,1]$, the following operator

$$D : v(x) \mapsto \begin{pmatrix} v'(x) \\ v(0) \end{pmatrix}$$

is a bijection between $V = C^1[0,1]$ and $W = C[0,1] \times \mathbb{R}$. □

If both V and W are normed spaces, we can talk about the continuity and boundedness of the operators.

Definition 2.1.6 *Let V and W be two normed spaces. An operator $T : V \to W$ is* continuous *at $v \in \mathcal{D}(T)$ if*

$$\{v_n\} \subset \mathcal{D}(T) \text{ and } v_n \to v \text{ in } V \quad \Longrightarrow \quad T(v_n) \to T(v) \text{ in } W.$$

T is said to be continuous *if it is continuous over its domain $\mathcal{D}(T)$. The operator is* bounded *if for any $r > 0$, there is an $R > 0$ such that*

$$v \in \mathcal{D}(T) \text{ and } \|v\| \le r \quad \Longrightarrow \quad \|T(v)\| \le R.$$

We observe that an alternative definition of the boundedness is that for any set $B \subset \mathcal{D}(T)$,

$$\sup_{v \in B} \|v\|_V < \infty \quad \Longrightarrow \quad \sup_{v \in B} \|T(v)\|_W < \infty.$$

Example 2.1.7 Let us consider the differentiation operator again. The spaces $C[0,1]$ and $C^1[0,1]$ are associated with their standard norms

$$\|v\|_{C[0,1]} = \max_{0 \le x \le 1} |v(x)|$$

and

$$\|v\|_{C^1[0,1]} = \|v\|_{C[0,1]} + \|v'\|_{C[0,1]}. \tag{2.1.2}$$

Then the operator

$$T_1 = \frac{d}{dx} : C^1[0,1] \subset C[0,1] \to C[0,1]$$

is not continuous using the infinity norm of $C[0,1]$ for $C^1[0,1]$, whereas the operator

$$T_2 = \frac{d}{dx} : C^1[0,1] \to C[0,1]$$

is continuous using the norm of (2.1.2) for $C^1[0,1]$. □

Exercise 2.1.1 Consider Example 2.1.7. Show that T_1 is unbounded and T_2 is bounded, as asserted in the example.

Exercise 2.1.2 Let $T_1 : C[a,b] \to C[a,b]$ be an operator defined by the formula

$$T_1 v(x) = \int_a^b (x - t)\, v(t)\, dt, \quad a \le x \le b, \ v \in C[a,b].$$

Determine the range of T_1. Is T_1 injective?

Exercise 2.1.3 Extending Exercise 2.1.2, for an arbitrary positive integer n, let $T_n : C[a,b] \to C[a,b]$ be defined by

$$T_n v(x) = \int_a^b (x-t)^n v(t)\,dt, \quad a \le x \le b,\ v \in C[a,b].$$

Determine the range of T_n and decide if T_n is injective.

Exercise 2.1.4 Over the space $C[a,b]$, define the operator T by

$$Tv(x) = \int_a^x v(t)\,dt, \quad a \le x \le b,\ v \in C[a,b].$$

Find the range of T. Is T a bijection between $C[a,b]$ and its range?

2.2 Continuous linear operators

This chapter is focused on the analysis of a particular type of operators called linear operators. From now on, when we write $T : V \to W$, we implicitly assume $\mathcal{D}(T) = V$, unless stated otherwise. As in Chapter 1, $\mathbb{K}$ denotes the set of scalars associated with the vector space under consideration.

Definition 2.2.1 *Let V and W be two linear spaces. An operator $L : V \to W$ is said to be* linear *if*

$$L(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1 L(v_1) + \alpha_2 L(v_2) \quad \forall\, v_1, v_2 \in V,\ \forall\, \alpha_1, \alpha_2 \in \mathbb{K}\,,$$

or equivalently,

$$\begin{aligned} L(v_1 + v_2) &= L(v_1) + L(v_2) \quad \forall\, v_1, v_2 \in V, \\ L(\alpha\, v) &= \alpha\, L(v) \quad \forall\, v \in V,\ \forall\, \alpha \in \mathbb{K}\,. \end{aligned}$$

For a linear operator L, we usually write $L(v)$ as Lv.

An important property of a linear operator is that continuity and boundedness are equivalent. We state and prove this result in the form of a theorem after two preparatory propositions which are themselves important.

Proposition 2.2.2 *Let V and W be normed spaces, $L : V \to W$ a linear operator. Then continuity of L over the whole space is equivalent to its continuity at any one point, say at $v = 0$.*

Proof. Continuity over the whole space certainly implies continuity at any point. Now assume L is continuous at $v = 0$:

$$v_n \to 0 \text{ in } V \implies Lv_n \to 0 \text{ in } W. \tag{2.2.1}$$

Let $v \in V$ be arbitrarily given and $\{v_n\} \subset V$ a sequence converging to v. Then $v_n - v \to 0$, and by (2.2.1), $L(v_n - v) = Lv_n - Lv \to 0$, i.e., $Lv_n \to Lv$. Hence L is continuous at v. □

Proposition 2.2.3 *Let V and W be normed spaces, $L : V \to W$ a linear operator. Then L is bounded if and only if there exists a constant $\gamma \geq 0$ such that*

$$\|Lv\|_W \leq \gamma \|v\|_V \quad \forall\, v \in V. \tag{2.2.2}$$

Proof. Obviously (2.2.2) implies the boundedness. Conversely, suppose L is bounded, then

$$\gamma \equiv \sup_{v \in B_1} \|Lv\|_W < \infty,$$

where $B_1 = \{v \in V \mid \|v\|_V \leq 1\}$ is the unit ball centered at 0. Now for any $v \neq 0$, $v/\|v\|_V \in B_1$ and by the linearity of L,

$$\|Lv\|_W = \|v\|_V \|L(v/\|v\|_V)\|_W \leq \gamma \, \|v\|_V,$$

i.e., (2.2.2) holds. □

Theorem 2.2.4 *Let V and W be normed spaces, $L : V \to W$ a linear operator. Then L is continuous on V if and only if it is bounded on V.*

Proof. Firstly we assume L is not bounded and prove that it is not continuous at 0. Since L is unbounded, we can find a bounded sequence $\{v_n\} \subset V$ such that $\|Lv_n\| \to \infty$. Without loss of generality, we may assume $Lv_n \neq 0$ for all n. Then we define a new sequence

$$\tilde{v}_n = \frac{v_n}{\|Lv_n\|_W}.$$

This sequence has the property that $\tilde{v}_n \to 0$ and $\|L\tilde{v}_n\|_W = 1$. Thus L is not continuous.

Secondly we assume L is bounded and show that it must be continuous. Indeed from (2.2.2) we have the Lipschitz inequality

$$\|Lv_1 - Lv_2\|_W \leq \gamma \, \|v_1 - v_2\|_V \quad \forall\, v_1, v_2 \in V, \tag{2.2.3}$$

which implies the continuity in an obvious fashion. □

From (2.2.3), we see that for a linear operator, continuity and Lipschitz continuity are equivalent.

We use the notation $\mathcal{L}(V, W)$ for the set of all the continuous linear operators from a normed space V to another normed space W. In the special case $W = V$, we use $\mathcal{L}(V)$ to replace $\mathcal{L}(V, V)$. We see that for

a linear operator, boundedness (2.2.2) is equivalent to continuity. Thus if $L \in \mathcal{L}(V, W)$, it is meaningful to define

$$\|L\|_{V,W} = \sup_{0 \neq v \in V} \frac{\|Lv\|_W}{\|v\|_V}. \tag{2.2.4}$$

Using the linearity of L, we have the following relations

$$\begin{aligned} \|L\|_{V,W} &= \sup_{v \in B_1} \|Lv\|_W = \sup_{v:\|v\|_V = 1} \|Lv\|_W \\ &= \frac{1}{r} \sup_{v:\|v\|_V = r} \|Lv\|_W = \frac{1}{r} \sup_{v:\|v\|_V \leq r} \|Lv\|_W \end{aligned}$$

for any $r > 0$. The norm $\|L\|_{V,W}$ is the maximum size in W of the image under L of the unit ball B_1 in V.

Theorem 2.2.5 *The set $\mathcal{L}(V, W)$ is a linear space, and* (2.2.4) *defines a norm over the space.*

We leave the proof of the theorem to the reader. The norm (2.2.4) is usually called the *operator norm* of L, which enjoys the following compatibility property

$$\|Lv\|_W \leq \|L\|_{V,W} \|v\|_V \quad \forall\, v \in V. \tag{2.2.5}$$

If it is not stated explicitly, we always understand the norm of an operator as an operator norm defined by (2.2.4). Another useful inequality involving operator norms is given in the following result.

Theorem 2.2.6 *Let U, V and W be normed spaces, $L_1 : U \to V$ and $L_2 : V \to W$ be continuous linear operators. Then the composite operator $L_2 L_1 : U \to W$ defined by*

$$L_2 L_1(v) = L_2(L_1(v)) \quad \forall\, v \in U$$

is a continuous linear mapping. Moreover,

$$\|L_2 L_1\|_{U,W} \leq \|L_1\|_{U,V} \|L_2\|_{V,W}. \tag{2.2.6}$$

Proof. The composite operator $L_2 L_1$ is obviously linear. We now prove (2.2.6), that also implies the continuity of $L_2 L_1$. By (2.2.5), for any $v \in U$,

$$\begin{aligned} \|L_2 L_1(v)\|_W &= \|L_2(L_1(v))\|_W \\ &\leq \|L_2\|_{V,W} \|L_1 v\|_V \\ &\leq \|L_2\|_{V,W} \|L_1\|_{U,V} \|v\|_U. \end{aligned}$$

Hence, (2.2.6) is valid. □

As an important special case, if V is a normed space and if $L \in \mathcal{L}(V)$, then for any non-negative integer n,

$$\|L^n\| \leq \|L\|^n.$$

The operator L^n is defined recursively: $L^n = L(L^{n-1})$, $L^n v = L(L^{n-1}v)$ $\forall\, v \in V$, and $L^0 = I$ is defined to be the identity operator. Both (2.2.5) and (2.2.6) are very useful relations for error analysis of some numerical methods.

For a linear operator, the null set $\mathcal{N}(L)$ becomes a subspace of V, and we have the statement

$$L \text{ is one-to-one} \quad \Longleftrightarrow \quad \mathcal{N}(L) = \{0\}.$$

Example 2.2.7 Let V be a linear space. Then the identity operator $I : V \to V$ belongs to $\mathcal{L}(V)$, and $\|I\| = 1$. □

Example 2.2.8 Recall Example 2.1.3. Let $V = \mathbb{C}^n$, $W = \mathbb{C}^m$, and $L(\boldsymbol{v}) = A\boldsymbol{v}$, $\boldsymbol{v} \in \mathbb{C}^n$, where $A = (a_{ij}) \in \mathbb{C}^{m\times n}$ is a complex matrix. If the norms on V and W are $\|\cdot\|_\infty$, then the operator norm is the matrix ∞-norm,

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|.$$

If the norms on V and W are $\|\cdot\|_1$, then the operator norm is the matrix 1-norm,

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|.$$

If the norms on V and W are $\|\cdot\|_2$, then the operator norm is the spectral norm

$$\|A\|_2 = \sqrt{r_\sigma(A^*A)} = \sqrt{r_\sigma(AA^*)},$$

where A^* denotes the conjugate transpose of A. For a square matrix B, $r_\sigma(B)$ denotes the spectral radius of the matrix B,

$$r_\sigma(B) = \max_{\lambda\in\sigma(B)} |\lambda|$$

and $\sigma(B)$ denotes the spectrum of B, the set of all the eigenvalues of B. Proofs of these results are given in [15, Section 7.3]. □

Example 2.2.9 Let $V = W = C[a,b]$ with the norm $\|\cdot\|_\infty$. Let $k \in C([a,b]^2)$, and define $K : C[a,b] \to C[a,b]$ by

$$(Kv)(x) = \int_a^b k(x,y)\, v(y)\, dy. \tag{2.2.7}$$

The mapping K in (2.2.7) is an example of a *linear integral operator*, and the function $k(\cdot,\cdot)$ is called the *kernel function* of the integral operator. Under the continuity assumption on $k(\cdot,\cdot)$, the integral operator is continuous from $C[a,b]$ to $C[a,b]$. Furthermore,

$$\|K\| = \max_{a\le x\le b} \int_a^b |k(x,y)|\, dy. \tag{2.2.8}$$

The linear integral operator (2.2.7) is later used extensively. □

2.2.1 $\mathcal{L}(V,W)$ as a Banach space

In approximating integral and differential equations, the integral or differential operator is often approximated by a sequence of operators of a simpler form. In such cases, it is important to consider the limits of convergent sequences of bounded operators, and this makes it important to have $\mathcal{L}(V,W)$ be a complete space.

Theorem 2.2.10 *Let V be a normed space, and W be a Banach space. Then $\mathcal{L}(V,W)$ is a Banach space.*

Proof. Let $\{L_n\}$ be a Cauchy sequence in $\mathcal{L}(V,W)$. This means

$$\epsilon_n \equiv \sup_{p\ge 1} \|L_{n+p} - L_n\| \to 0 \quad \text{as } n \to \infty.$$

We must define a limit for $\{L_n\}$ and show that it belongs to $\mathcal{L}(V,W)$.

For each $v \in V$,

$$\|L_{n+p}v - L_n v\|_W \le \epsilon_n \|v\|_V \to 0 \quad \text{as } n \to \infty. \tag{2.2.9}$$

Thus $\{L_n v\}$ is a Cauchy sequence in W. Since W is complete, the sequence has a limit, denoted by $L(v)$. This defines an operator $L : V \to W$. Let us prove that L is linear, bounded, and $\|L_n - L\|_{V,W} \to 0$ as $n \to \infty$.

For any $v_1, v_2 \in V$ and $\alpha_1, \alpha_2 \in \mathbb{K}$,

$$\begin{aligned} L(\alpha_1 v_1 + \alpha_2 v_2) &= \lim_{n\to\infty} L_n(\alpha_1 v_1 + \alpha_2 v_2) \\ &= \lim_{n\to\infty} (\alpha_1 L_n v_1 + \alpha_2 L_n v_2) \\ &= \alpha_1 \lim_{n\to\infty} L_n v_1 + \alpha_2 \lim_{n\to\infty} L_n v_2 \\ &= \alpha_1 L(v_1) + \alpha_2 L(v_2). \end{aligned}$$

Thus L is linear.

Now for any $v \in V$, we take the limit $p \to \infty$ in (2.2.9) to obtain

$$\|Lv - L_n v\|_W \le \epsilon_n \|v\|_V.$$

Thus

$$\|L - L_n\|_{V,W} = \sup_{\|v\|_V \le 1} \|Lv - L_n v\|_W \le \epsilon_n \to 0 \quad \text{as } n \to \infty.$$

Hence $L \in \mathcal{L}(V,W)$ and $L_n \to L$ as $n \to \infty$. □

Exercise 2.2.1 For a linear operator $L : V \to W$, show that $L(0) = 0$.

Exercise 2.2.2 Prove that if a linear operator is discontinuous at one point, then it is discontinuous everywhere.

Exercise 2.2.3 Prove that $\mathcal{L}(V,W)$ is a linear space, and (2.2.4) defines a norm on the space.

Exercise 2.2.4 For $L \in \mathcal{L}(V,W)$, show that its norm is given by

$$\|L\| = \inf\{\alpha \mid \|Lv\|_W \le \alpha \|v\|_V \ \forall\, v \in V\}.$$

Exercise 2.2.5 Assume $k \in C([a,b]^2)$. Show that the integral operator K defined by (2.2.7) is continuous, and its operator norm is given by the formula (2.2.8).

Exercise 2.2.6 Let $\Omega \subset \mathbb{R}^d$ be a domain, $1 \le p \le \infty$. Given $m \in C(\overline{\Omega})$, define an operator $M : L^p(\Omega) \to L^p(\Omega)$ by the formula

$$Mv(\boldsymbol{x}) = m(\boldsymbol{x})\, v(\boldsymbol{x}), \quad v \in L^p(\Omega).$$

Show that M is linear, bounded, and $\|M\| = \|m\|_{C(\overline{\Omega})}$.

Exercise 2.2.7 A linear operator L is called nonsingular if $\mathcal{N}(L) = \{0\}$. Otherwise, it is called singular. Show that if L is nonsingular, then a solution of the equation $Lu = f$ is unique.

Exercise 2.2.8 Let a linear operator $L : V \to W$ be nonsingular and map V onto W. Show that for each $f \in W$, the equation $Lu = f$ has a unique solution $u \in V$.

2.3 The geometric series theorem and its variants

The following result is used commonly in numerical analysis and applied mathematics. It is also the means by which we can analyze the solvability of problems that are "close" to another problem known to be uniquely solvable.

Theorem 2.3.1 (Geometric series theorem) *Let V be a Banach space, $L \in \mathcal{L}(V)$. Assume*

$$\|L\| < 1. \tag{2.3.1}$$

Then $I - L$ is a bijection on V, its inverse is a bounded linear operator,

$$(I - L)^{-1} = \sum_{n=0}^{\infty} L^n,$$

and

$$\|(I - L)^{-1}\| \leq \frac{1}{1 - \|L\|}. \tag{2.3.2}$$

Proof. Define a sequence in $\mathcal{L}(V)$: $M_n = \sum_{i=0}^{n} L^i$, $n \geq 0$. For $p \geq 1$,

$$\|M_{n+p} - M_n\| = \left\| \sum_{i=n+1}^{n+p} L^i \right\| \leq \sum_{i=n+1}^{n+p} \|L^i\| \leq \sum_{i=n+1}^{n+p} \|L\|^i.$$

Using the assumption (2.3.1), we have

$$\|M_{n+p} - M_n\| \leq \frac{\|L\|^{n+1}}{1 - \|L\|}. \tag{2.3.3}$$

Hence,

$$\sup_{p \geq 1} \|M_{n+p} - M_n\| \to 0 \quad \text{as } n \to \infty,$$

and $\{M_n\}$ is a Cauchy sequence in $\mathcal{L}(V)$. Since $\mathcal{L}(V)$ is complete, there is an $M \in \mathcal{L}(V)$ with

$$\|M_n - M\| \to 0 \quad \text{as } n \to \infty.$$

Using the definition of M_n and simple algebraic manipulation,

$$(I - L)\, M_n = M_n (I - L) = I - L^{n+1}.$$

Let $n \to \infty$ to get

$$(I - L)\, M = M\, (I - L) = I.$$

This relation implies $(I - L)$ is a bijection, and

$$M = (I - L)^{-1} = \lim_{n \to \infty} \sum_{i=0}^{n} L^i = \sum_{n=0}^{\infty} L^n.$$

To prove the bound (2.3.2), note that

$$\|M_n\| \leq \sum_{i=0}^{n} \|L\|^i \leq \frac{1}{1 - \|L\|}.$$

Taking the limit $n \to \infty$, we obtain (2.3.2). □

The theorem says that under the stated assumptions, for any $f \in V$, the equation

$$(I - L)\, u = f \tag{2.3.4}$$

has a unique solution $u = (I - L)^{-1} f \in V$. Moreover, the solution depends continuously on the right hand side f: Letting $(I - L)\, u_1 = f_1$ and $(I - L)\, u_2 = f_2$, it follows that

$$u_1 - u_2 = (I - L)^{-1}(f_1 - f_2),$$

and so

$$\|u_1 - u_2\| \le c\, \|f_1 - f_2\|$$

with $c = 1/(1 - \|L\|)$.

The theorem also provides an approach to approximate the solution of the equation (2.3.4). Under the stated assumptions of the theorem, we have

$$u = \lim_{n \to \infty} u_n$$

where

$$u_n = \sum_{j=0}^{n} L^j f. \tag{2.3.5}$$

Example 2.3.2 Consider the linear integral equation of the second kind

$$\lambda\, u(x) - \int_a^b k(x, y)\, u(y)\, dy = f(x), \quad a \le x \le b \tag{2.3.6}$$

with $\lambda \neq 0$, $k(x, y)$ continuous for $x, y \in [a, b]$, and $f \in C[a, b]$. Let $V = C[a, b]$ with the norm $\|\cdot\|_\infty$. Symbolically, we write the equation (2.3.6) as

$$(\lambda\, I - K)\, u = f, \tag{2.3.7}$$

where K is the linear integral operator generated by the kernel function $k(\cdot, \cdot)$. We also will often write this as $(\lambda - K)\, u = f$, understanding it to mean the same as in (2.3.7).

This equation (2.3.7) can be converted into the form needed in the geometric series theorem:

$$(I - L)\, u = \frac{1}{\lambda} f, \qquad L = \frac{1}{\lambda} K.$$

Applying the geometric series theorem we assert that if

$$\|L\| = \frac{1}{|\lambda|} \|K\| < 1,$$

then $(I-L)^{-1}$ exists and

$$\|(I-L)^{-1}\| \le \frac{1}{1-\|L\|}.$$

Equivalently, if

$$\|K\| = \max_{a\le x\le b} \int_a^b |K(x,y)|\,dy < |\lambda|, \tag{2.3.8}$$

then $(\lambda I - K)^{-1}$ exists and

$$\|(\lambda I-K)^{-1}\| \le \frac{1}{|\lambda| - \|K\|}.$$

Hence under the assumption (2.3.8), for any $f \in C[a,b]$, the integral equation (2.3.6) has a unique solution $u \in C[a,b]$ and

$$\|u\|_\infty \le \|(\lambda I-K)^{-1}\|\,\|f\|_\infty \le \frac{\|f\|_\infty}{|\lambda| - \|K\|}. \qquad \square$$

We observe that the geometric series theorem is a straightforward generalization to linear continuous operators on a Banach space of the power series

$$(1-x)^{-1} = \sum_{n=0}^{\infty} x^n, \quad x \in \mathbb{R},\ |x| < 1$$

or its complex version

$$(1-z)^{-1} = \sum_{n=0}^{\infty} z^n, \quad z \in \mathbb{C},\ |z| < 1.$$

From the proof of the theorem we see that for a linear operator $L \in \mathcal{L}(V)$ over a Banach space V, if $\|L\| < 1$, then the series $\sum_{n=0}^{\infty} L^n$ converges in $\mathcal{L}(V)$ and the value of the series is the operator $(I-L)^{-1}$. More generally, we can similarly define an operator-valued function $f(L)$ of an operator variable L from a real function $f(x)$ of a real variable x (or a complex-valued function $f(z)$ of a complex variable z), as long as $f(x)$ is analytic at $x = 0$, i.e., $f(x)$ has a convergent power series expansion:

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \quad |x| < \gamma$$

for some constant $\gamma > 0$, where $a_n = f^{(n)}(0)/n!$, $n \ge 0$. Now if V is a Banach space and $L \in \mathcal{L}(V)$ satisfies $\|L\| < \gamma$, then we define

$$f(L) = \sum_{n=0}^{\infty} a_n L^n.$$

The series on the right-hand side is a well-defined operator in $\mathcal{L}(V)$, thanks to the assumption $\|L\| < \gamma$. We now give some examples of operator-valued functions obtained by this approach, with $L \in \mathcal{L}(V)$ and V a Banach space:

$$e^L = \sum_{n=0}^{\infty} \frac{1}{n!} L^n,$$
$$\sin(L) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} L^{2n+1},$$
$$\arctan(L) = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} L^{2n+1}, \quad \|L\| < 1.$$

2.3.1 A generalization

To motivate a generalization of Theorem 2.3.1, consider the *Volterra integral equation of the second kind*

$$u(x) - \int_0^x \ell(x,y)u(y)\,dy = f(x), \quad x \in [0,B]. \tag{2.3.9}$$

Here, $B > 0$ and we assume the kernel function $\ell(x,y)$ is continuous for $0 \le y \le x \le B$, and $f \in C[0,B]$. If we apply Theorem 2.3.1, then we will need to assume a condition such as

$$\max_{x\in[0,B]} \int_0^x |\ell(x,y)|\,dy < 1$$

in order to conclude the unique solvability. However, we can use a variant of the geometric series theorem to show that the equation (2.3.9) is uniquely solvable, irregardless of the size of the kernel function $\ell(x,y)$. Symbolically, we write this integral equation as $(I-L)u = f$.

Corollary 2.3.3 *Let V be a Banach space, $L \in \mathcal{L}(V)$. Assume for some integer $m \ge 1$ that*

$$\|L^m\| < 1. \tag{2.3.10}$$

Then $I - L$ is a bijection on V, its inverse is a bounded linear operator, and

$$\|(I-L)^{-1}\| \le \frac{1}{1-\|L^m\|} \sum_{i=0}^{m-1} \|L^i\|. \tag{2.3.11}$$

Proof. From Theorem 2.3.1, we know that $(I-L^m)^{-1}$ exists as a bounded bijective operator on V to V,

$$(I-L^m)^{-1} = \sum_{j=0}^{\infty} L^{mj} \quad \text{in } \mathcal{L}(V)$$

and

$$\|(I - L^m)^{-1}\| \leq \frac{1}{1 - \|L^m\|}.$$

From the identities

$$(I - L)\left(\sum_{i=0}^{m-1} L^i\right) = \left(\sum_{i=0}^{m-1} L^i\right)(I - L) = I - L^m,$$

we then conclude that $(I - L)$ is a bijection,

$$(I - L)^{-1} = \left(\sum_{i=0}^{m-1} L^i\right)(I - L^m)^{-1}$$

and the bound (2.3.11) holds. □

Example 2.3.4 Returning to (2.3.9), define

$$Lv(x) = \int_0^x \ell(x, y)v(y)\, dy, \quad 0 \leq x \leq B,\ v \in C[0, B].$$

Easily, L is a bounded linear operator on $C[0, B]$ to itself. The iterated operators L^k take the form

$$L^k v(x) = \int_0^x \ell_k(x, y)v(y)\, dy$$

for $k = 2, 3, \ldots$, and $\ell_1(x, y) \equiv \ell(x, y)$. It is straightforward to show

$$\ell_{k+1}(x, y) = \int_y^x \ell_k(x, z)\ell(z, y)\, dz, \quad k = 1, 2, \ldots.$$

Let

$$M = \max_{0 \leq y \leq x \leq B} |\ell(x, y)|.$$

Then

$$|\ell_1(x, y)| \leq M, \quad 0 \leq y \leq x \leq B.$$

Assume for an integer $k \geq 1$,

$$|\ell_k(x, y)| \leq M^k \frac{(x - y)^{k-1}}{(k - 1)!}, \quad 0 \leq y \leq x \leq B.$$

Then for $0 \leq y \leq x \leq B$,

$$\begin{aligned} |\ell_{k+1}(x, y)| &\leq \int_y^x |\ell_k(x, z)|\, |\ell(z, y)|\, dz \\ &\leq M^{k+1} \int_y^x \frac{(x - z)^{k-1}}{(k - 1)!}\, dz \\ &= M^{k+1} \frac{(x - y)^k}{k!}. \end{aligned}$$

Hence, for any integer $k \geq 1$ and any $x \in [0, B]$,

$$\begin{aligned} |L^k v(x)| &\leq \int_0^x |\ell_k(x,y)|\,|v(y)|\,dy \\ &\leq \int_0^x M^k \frac{(x-y)^{k-1}}{(k-1)!}\,dy\,\|v\|_\infty \\ &= \frac{M^k x^k}{k!}\,\|v\|_\infty, \end{aligned}$$

and then

$$\|L^k\| \leq \frac{M^k B^k}{k!}, \quad k = 1, 2, \ldots.$$

It is clear that the right side converges to zero as $k \to \infty$, and thus (2.3.10) is satisfied for m large enough. We can also use this result to construct bounds for the solutions of (2.3.9), which we leave to Exercise 2.3.11. □

2.3.2 A perturbation result

An important technique in applied mathematics is to study an equation by relating it to a "nearby" equation for which there is a known solvability result. One of the more popular tools is the following perturbation theorem.

Theorem 2.3.5 *Let V and W be normed spaces with at least one of them being complete. Assume $L \in \mathcal{L}(V, W)$ has a bounded inverse $L^{-1} : W \to V$. Assume $M \in \mathcal{L}(V, W)$ satisfies*

$$\|M - L\| < \frac{1}{\|L^{-1}\|}. \tag{2.3.12}$$

Then $M : V \to W$ is a bijection, $M^{-1} \in \mathcal{L}(W, V)$ and

$$\|M^{-1}\| \leq \frac{\|L^{-1}\|}{1 - \|L^{-1}\|\,\|L - M\|}. \tag{2.3.13}$$

Moreover,

$$\|L^{-1} - M^{-1}\| \leq \frac{\|L^{-1}\|^2 \|L - M\|}{1 - \|L^{-1}\|\,\|L - M\|}. \tag{2.3.14}$$

For solutions of the equations $Lv_1 = w$ and $Mv_2 = w$, we have the bound

$$\|v_1 - v_2\| \leq \|M^{-1}\|\,\|(L - M)\,v_1\|. \tag{2.3.15}$$

Proof. We write M as a perturbation of L. If W is complete, we write

$$M = [I - (L - M)\,L^{-1}]\,L;$$

whereas if V is complete, we write

$$M = L\,[I - L^{-1}(L - M)].$$

Let us prove the result for the case W is complete.

The operator $(L - M)\,L^{-1} \in \mathcal{L}(W)$ satisfies

$$\|(L - M)\,L^{-1}\| \leq \|L - M\|\,\|L^{-1}\| < 1.$$

Thus by the geometric series theorem, $[I - (L - M)\,L^{-1}]^{-1}$ exists and

$$\|[I - (L - M)\,L^{-1}]^{-1}\| \leq \frac{1}{1 - \|(L - M)\,L^{-1}\|} \leq \frac{1}{1 - \|L^{-1}\|\,\|L - M\|}.$$

So M^{-1} exists with

$$M^{-1} = L^{-1}[I - (L - M)\,L^{-1}]^{-1}$$

and

$$\|M^{-1}\| \leq \|L^{-1}\|\,\|[I - (L - M)\,L^{-1}]^{-1}\| \leq \frac{\|L^{-1}\|}{1 - \|L^{-1}\|\,\|L - M\|}.$$

To prove (2.3.14), we write

$$L^{-1} - M^{-1} = M^{-1}(M - L)\,L^{-1},$$

take norms and use (2.3.13). For (2.3.15), write

$$v_1 - v_2 = (L^{-1} - M^{-1})\,w = M^{-1}(M - L)\,L^{-1}w = M^{-1}(M - L)\,v_1$$

and take norms and bounds. □

The above theorem can be paraphrased as follows: *An operator that is close to an operator with a bounded inverse will itself have a bounded inverse.* This is the framework for innumerable solvability results for linear differential and integral equations, and variations of it are also used with nonlinear operator equations.

The inequality (2.3.14) can be termed the local Lipschitz continuity of the operator inverse. The bound (2.3.15) can be used both as an *a priori* and an *a posteriori* error estimate, depending on the way we use it. First, let us view the equation $Lv = w$ as the exact problem, and we take a sequence of approximation problems $L_n v_n = w$, $n = 1, 2, \ldots$. Assuming the sequence $\{L_n\}$ converges to L, we can apply the perturbation theorem to conclude that at least for sufficiently large n, the equation $L_n v_n = w$ has a unique solution v_n, and we have the error estimate

$$\|v - v_n\| \leq \|L_n^{-1}\|\,\|(L - L_n)\,v\|. \tag{2.3.16}$$

The *consistency* of the approximation is defined by the condition

$$\|(L - L_n)\, v\| \to 0,$$

whereas the *stability* is defined by the condition that $\{\|L_n^{-1}\|\}_{n \text{ large}}$ is uniformly bounded. We see that consistency plus stability implies *convergence*:

$$\|v - v_n\| \to 0.$$

The error estimate (2.3.16) provides sufficient conditions for convergence (and order error estimate under regularity assumptions on the solution v) before we actually solve the approximation problem $L_n v_n = w$. Such an estimate is called an *a priori* error estimate. We notice that usually an a priori error estimate does not tell us quantitatively how small is the error.

Another way to use (2.3.15) is to view $Mv = w$ as the exact problem, and $L = M_n$ an approximation of M, $n = 1, 2, \ldots$. Denote by v_n the solution of the approximation equation $M_n v_n = w$; the equation is uniquely solvable at least for sufficiently large n. Then we have the error estimate

$$\|v - v_n\| \le \|M^{-1}\|\, \|(M - M_n)\, v_n\|.$$

Suppose we can estimate the term $\|M^{-1}\|$. Then after the approximate solution v_n is found, the above estimate offers a numerical upper bound for the error. Such an estimate is called an *a posteriori* error estimate.

Example 2.3.6 We examine the solvability of the integral equation

$$\lambda\, u(x) - \int_0^1 \sin(xy)\, u(y)\, dy = f(x), \quad 0 \le x \le 1 \tag{2.3.17}$$

with $\lambda \neq 0$. From the discussion of the Example 2.3.2, if

$$|\lambda| > \|K\| = \int_0^1 \sin(y)\, dy = 1 - \cos(1) \approx 0.4597, \tag{2.3.18}$$

then for every $f \in C[0,1]$, (2.3.17) admits a unique solution $u \in C[0,1]$.

To extend the values of λ for which (2.3.17) has a unique solution, we apply the perturbation theorem. Since $\sin(xy) \approx xy$ for small values of $|xy|$, we compare (2.3.17) with

$$\lambda\, v(x) - \int_0^1 x\, y\, v(y)\, dy = f(x), \quad 0 \le x \le 1. \tag{2.3.19}$$

In the notation of the perturbation theorem, equation (2.3.17) is $Mu = f$ and (2.3.19) is $Lv = f$. The normed space is $V = C[0,1]$ with the norm $\|\cdot\|_\infty$, and $L, M \in \mathcal{L}(V)$.

The integral equation (2.3.19) can be solved explicitly. From (2.3.19), assuming $\lambda \neq 0$, we have that every solution v takes the form

$$v(x) = \frac{1}{\lambda}\left[f(x) + cx\right]$$

for some constant c. Substituting this back into (2.3.19) leads to a formula for c, and then

$$v(x) = \frac{1}{\lambda}\left[f(x) + \frac{1}{\lambda - 1/3}\int_0^1 x\,y\,f(y)\,dy\right] \quad \text{if } \lambda \neq 0, \frac{1}{3}. \tag{2.3.20}$$

The relation (2.3.20) defines $L^{-1}f$ for all $f \in C[0,1]$.

To use the perturbation theorem, we need to measure several quantities. It can be computed that

$$\|L^{-1}\| \leq \frac{1}{|\lambda|}\left(1 + \frac{1}{2\,|\lambda - 1/3|}\right)$$

and

$$\|L - M\| = \int_0^1 (y - \sin y)\,dy = \cos(1) - \frac{1}{2} \approx 0.0403.$$

The condition (2.3.12) is implied by

$$\frac{1}{|\lambda|}\left(1 + \frac{1}{2\,|\lambda - 1/3|}\right) < \frac{1}{\cos(1) - 1/2}. \tag{2.3.21}$$

A graph of the left side of this inequality is given in Figure 2.1. If λ is assumed to be real, then there are three cases to be considered: $\lambda > 1/3$, $0 < \lambda < 1/3$, and $\lambda < 0$. For the case $\lambda < 0$, (2.3.21) is true if and only if $\lambda < \lambda_0 \approx -0.0881$, the negative root of the equation

$$\lambda^2 - \left(\frac{5}{6} - \cos 1\right)\lambda - \frac{5}{6}\left(\cos 1 - \frac{1}{2}\right) = 0.$$

As a consequence of the perturbation theorem, we have that if $\lambda < \lambda_0$, then (2.3.17) is uniquely solvable for all $f \in C[0,1]$. This is a significant improvement over the negative portion of the condition (2.3.18). Bounds can also be given on the solution u, but these are left to the reader, as are the remaining two cases for λ. □

Exercise 2.3.1 Consider the integral equation

$$\lambda u(x) - \int_0^1 \frac{u(y)}{1 + x^2y^2}\,dy = f(x), \quad 0 \leq x \leq 1$$

for a given $f \in C[0,1]$. Show this equation has a unique continuous solution u if $|\lambda|$ is chosen sufficiently large. For such values of λ, bound the solution u in terms of $\|f\|_\infty$.

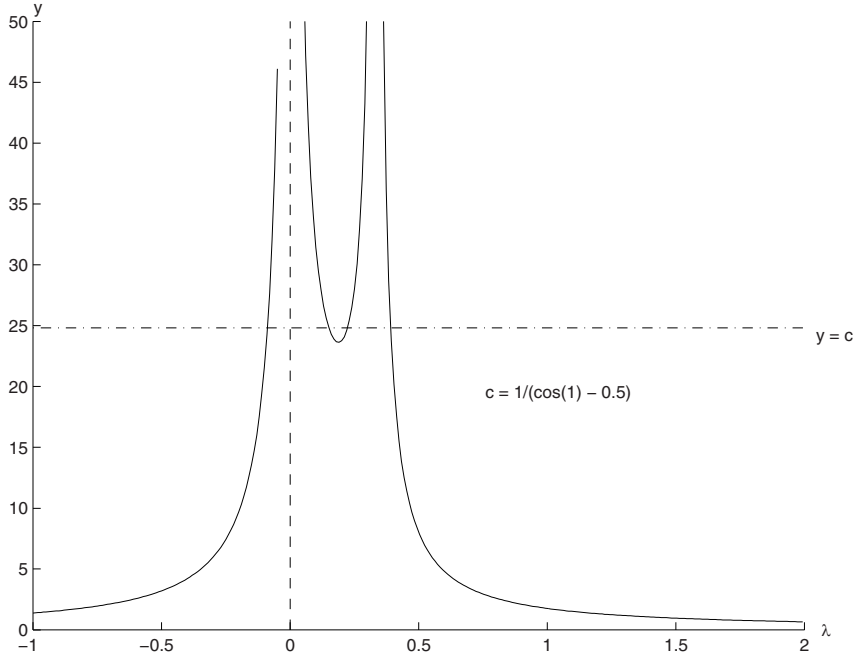

FIGURE 2.1. Graph of the left-hand side of inequality (2.3.21)

Exercise 2.3.2 Show that the integral equation

$$u(x) - \int_0^1 \sin \pi(x-t)\, u(t)\, dt = f(x)$$

has a unique solution $u \in C[0,1]$ for any given $f \in C[0,1]$. As an approximation of the solution u, use the formula (2.3.5) to compute u_2.

Exercise 2.3.3 Let V and W be Banach spaces. Assume $L \in \mathcal{L}(V,W)$ has a bounded inverse $L^{-1} : W \to V$ and $\{M_\lambda\}_{\lambda\in\Lambda} \subset \mathcal{L}(V,W)$ is a family of linear continuous operators such that $\|M_\lambda\| \le \varepsilon$, $\forall\, \lambda \in \Lambda$. Define $L_\lambda = L + M_\lambda$. Find a condition on ε so that given $f \in W$, for any $\lambda \in \Lambda$, the equation $L_\lambda u_\lambda = f$ has a unique solution $u_\lambda \in V$, and the iteration method: $u_0 \in V$ chosen, and for $n = 0, 1, \ldots,$

$$u_{n+1} = L^{-1}(f - M_\lambda u_n),$$

converges to the solution u_λ.

Exercise 2.3.4 A linear continuous operator L is said to be nilpotent if for some integer $m \ge 1$, $L^m = 0$. Show that if L is nilpotent, then $(I-L)^{-1}$ exists. Find a formula for $(I-L)^{-1}$.

Exercise 2.3.5 Complete the solvability analysis for Example 2.3.6.

Exercise 2.3.6 Repeat the solvability analysis of Example 2.3.6 for the integral equation

$$\lambda\, u(x) - \int_0^1 u(y) \arctan(xy)\, dy = f(x), \quad 0 \le x \le 1.$$

Use the approximation based on the Taylor approximation

$$\arctan(s) \approx s$$

for small values of s.

Exercise 2.3.7 Let V be a Banach space, $L \in \mathcal{L}(V)$. Assume $\|I - L\| < 1$. Show that L has a bounded inverse, and

$$L^{-1} = \sum_{i=0}^{\infty} (I - L)^i.$$

Exercise 2.3.8 Assume the conditions of the geometric series theorem are satisfied. Then for any $f \in V$, the equation $(I - L)\, u = f$ has a unique solution $u \in V$. Show that this solution can be approximated by a sequence $\{u_n\}$ defined by: $u_0 \in V$, $u_n = f + Lu_{n-1}$, $n = 1, 2, \ldots$. Derive an error bound for $\|u - u_n\|$.

Exercise 2.3.9 Let $f \in C[0, 1]$. Show that the continuous solution of the boundary value problem

$$\begin{aligned} -u''(x) &= f(x), \quad 0 < x < 1, \\ u(0) &= u(1) = 0 \end{aligned}$$

is

$$u(x) = \int_0^1 k(x, y)\, f(y)\, dy,$$

where the kernel function $k(x, y) = \min(x, y)\,(1 - \max(x, y))$. Let $a \in C[0, 1]$. Apply the geometric series theorem to show that the boundary value problem

$$\begin{aligned} -u''(x) + a(x)\, u(x) &= f(x), \quad 0 < x < 1, \\ u(0) &= u(1) = 0 \end{aligned}$$

has a unique continuous solution u if $\max_{0 \le x \le 1} |a(x)| \le a_0$ is sufficiently small. Give an estimate of the value a_0.

Exercise 2.3.10 Let V and W be Banach spaces. Assume $L \in \mathcal{L}(V, W)$ has a bounded inverse $L^{-1} : W \to V$ and $M \in \mathcal{L}(V, W)$ satisfies

$$\|M - L\| < \frac{1}{2\,\|L^{-1}\|}.$$

For any $f \in W$, the equation $Lu = f$ has a unique solution $u \in V$ which is approximated by the following iteration method: choose an initial guess $u_0 \in V$ and define

$$u_{n+1} = u_n + M^{-1}(f - Lu_n), \quad n \ge 0.$$

Prove the convergence $u_n \to u$ as $n \to \infty$.

This iteration method is useful where it is much easier to compute $M^{-1}g$ than $L^{-1}g$ for $g \in W$.

Exercise 2.3.11 Recall the Volterra equation (2.3.9). Bound the solution u using Corollary 2.3.3. Separately, obtain a bound for u by examining directly the convergence of the series

$$u = \sum_{k=0}^{\infty} L^k f$$

and relating it to the Taylor series for $\exp(M\,B)$.

2.4 Some more results on linear operators

In this section, we collect together several independent results which are important in working with linear operators.

2.4.1 An extension theorem

Bounded operators are often defined on a subspace of a larger space, and it is desirable to extend the domain of the original operator to the larger space, while retaining the boundedness of the operator.

Theorem 2.4.1 (EXTENSION THEOREM) *Let V be a normed space, and let $\widehat{V}$ denote its completion. Let W be a Banach space. Assume $L \in \mathcal{L}(V, W)$. Then there is a unique operator $\widehat{L} \in \mathcal{L}(\widehat{V}, W)$ with*

$$\widehat{L}v = L\,v \quad \forall\, v \in V$$

and

$$\|\widehat{L}\|_{\widehat{V},W} = \|L\|_{V,W}.$$

The operator $\widehat{L}$ is called an extension *of L.*

Proof. Given $v \in \widehat{V}$, let $\{v_n\} \subset V$ with $v_n \to v$ in $\widehat{V}$. The sequence $\{Lv_n\}$ is a Cauchy sequence in W by the following inequality

$$\|Lv_{n+p} - Lv_n\| \le \|L\|\,\|v_{n+p} - v_n\|.$$

Since W is complete, there is a limit $\widehat{L}(v) \in W$. We must show that $\widehat{L}$ is well-defined (i.e., $\widehat{L}(v)$ does not depend on the choice of the sequence $\{v_n\}$), linear and bounded.

To show $\widehat{L}$ is well-defined, let $v_n \to v$ and $\tilde{v}_n \to v$ with $\{v_n\}, \{\tilde{v}_n\} \subset V$. Then as $n \to \infty$,

$$\|Lv_n - L\tilde{v}_n\| \le \|L\|\,\|v_n - \tilde{v}_n\| \le \|L\|\,(\|v_n - v\| + \|\tilde{v}_n - v\|) \to 0.$$

Thus $\{Lv_n\}$ and $\{L\tilde{v}_n\}$ must have the same limit.

To show the linearity, let $u_n \to u$ and $v_n \to v$, and let $\alpha, \beta \in \mathbb{K}$. Then

$$\widehat{L}(\alpha\, u + \beta\, v) = \lim_{n\to\infty} L(\alpha\, u_n + \beta\, v_n) = \lim_{n\to\infty} (\alpha\, L u_n + \beta\, L v_n) = \alpha\, \widehat{L} u + \beta\, \widehat{L} v.$$

To show the boundedness, let $v_n \to v$ and $\{v_n\} \subset V$. Then taking the limit $n \to \infty$ in

$$\|Lv_n\|_W \le \|L\|\, \|v_n\|_V = \|L\|\, \|v_n\|_{\widehat{V}},$$

we obtain

$$\|\widehat{L} v\|_W \le \|L\|\, \|v\|_{\widehat{V}}.$$

So $\widehat{L}$ is bounded and

$$\|\widehat{L}\| = \sup_{0\neq v\in\widehat{V}} \frac{\|\widehat{L} v\|_W}{\|v\|_{\widehat{V}}} \le \|L\|.$$

To see that $\|\widehat{L}\|_{\widehat{V},W} = \|L\|_{V,W}$, we note

$$\|L\|_{V,W} = \sup_{0\neq v\in V} \frac{\|Lv\|_W}{\|v\|_V} = \sup_{0\neq v\in V} \frac{\|\widehat{L} v\|_W}{\|v\|_{\widehat{V}}} \le \sup_{0\neq v\in\widehat{V}} \frac{\|\widehat{L} v\|_W}{\|v\|_{\widehat{V}}} = \|\widehat{L}\|_{\widehat{V},W}.$$

To show that $\widehat{L}$ is unique, let $\widetilde{L}$ be another extension of L to $\widehat{V}$. Let $v \in \widehat{V}$ and let $v_n \to v$, $\{v_n\} \subset V$. Then

$$\|\widetilde{L} v - L v_n\|_W = \|\widetilde{L} v - \widetilde{L} v_n\|_W \le \|\widetilde{L}\|\, \|v - v_n\| \to 0 \quad \text{as } n \to \infty.$$

This shows $Lv_n \to \widetilde{L} v$ as $n \to \infty$. On the other hand, $Lv_n \to \widehat{L} v$. So we must have $\widetilde{L} v = \widehat{L} v$, for any $v \in \widehat{V}$. Therefore, $\widetilde{L} = \widehat{L}$. □

There are a number of ways in which this theorem can be used. Often we wish to work with linear operators which are defined and bounded on some normed space, but the space is not complete with the given norm. Since most function space arguments require complete spaces, the above theorem allows us to proceed with our arguments on a larger complete space, with an operator which agrees with our original one on the original space.

Example 2.4.2 Let $V = C^1[0,1]$ with the inner product norm

$$\|v\|_{1,2} = (\|v\|_2^2 + \|v'\|_2^2)^{1/2}.$$

The completion of $C^1[0,1]$ with respect to $\|\cdot\|_{1,2}$ is the Sobolev space $H^1(0,1)$, which was introduced earlier in Example 1.3.7. (Details of Sobolev spaces are given later in Chapter 7.) Let $W = L^2(0,1)$ with the standard norm $\|\cdot\|_2$.

Define the differentiation operator $D : C^1[0,1] \to L^2(0,1)$ by

$$(Dv)(x) = v'(x), \quad 0 \le x \le 1,\ v \in C^1[0,1].$$

We have

$$\|Dv\|_2 = \|v'\|_2 \leq \|v\|_{1,2},$$

and thus

$$\|D\|_{V,W} \leq 1.$$

By the extension theorem, we can extend D to $\widehat{D} \in \mathcal{L}(H^1(0,1), L^2(0,1))$, the differentiation operator on $H^1(0,1)$. A more concrete realization of $\widehat{D}$ can be obtained using the notion of weak derivatives. This is also discussed in Chapter 7. □

2.4.2 Open mapping theorem

This theorem is widely used in obtaining boundedness of inverse operators. When considered in the context of solving an equation $Lv = w$, the theorem says that *existence* and *uniqueness* of solutions for all $w \in W$ implies the *stability* of the solution v, i.e., "small changes" in the given data w causes only "small changes" in the solution v. For a proof of this theorem, see [58, p. 91] or [250, p. 179].

Theorem 2.4.3 *Let V and W be Banach spaces. If $L \in \mathcal{L}(V,W)$ is a bijection, then $L^{-1} \in \mathcal{L}(W,V)$.*

To be more precise concerning the stability of the problem being solved, let $Lv = w$ and $L\hat{v} = \hat{w}$. We then have

$$v - \hat{v} = L^{-1}(w - \hat{w}),$$

and then

$$\|v - \hat{v}\| \leq \|L^{-1}\| \, \|w - \hat{w}\|.$$

As $w - \hat{w}$ becomes small, so must $v - \hat{v}$. The term $\|L^{-1}\|$ gives a relationship between the size of the error in the data w and that of the error in the solution v. A more important way is to consider the relative changes in the two errors:

$$\frac{\|v - \hat{v}\|}{\|v\|} \leq \frac{\|L^{-1}\| \, \|w - \hat{w}\|}{\|v\|} = \|L^{-1}\| \|L\| \, \frac{\|w - \hat{w}\|}{\|L\| \, \|v\|}.$$

Applying $\|w\| \leq \|L\| \, \|v\|$, we obtain

$$\frac{\|v - \hat{v}\|}{\|v\|} \leq \|L^{-1}\| \, \|L\| \frac{\|w - \hat{w}\|}{\|w\|}. \tag{2.4.1}$$

The quantity $\text{cond}(L) \equiv \|L^{-1}\| \, \|L\|$ is called the *condition number* of the equation, and it relates the relative errors in the data w and in the solution v. Note that we always have $\text{cond}(L) \geq 1$ as

$$\|L^{-1}\| \, \|L\| \geq \|L^{-1}L\| = \|I\| = 1.$$

Problems with a small condition number are called *well-conditioned*, whereas those with a large condition number *ill-conditioned*.

In a related vein, consider a problem $Lv = w$, $L : V \to W$, in which L is bounded and injective, but not surjective. The inverse operator L^{-1} exists on the range $\mathcal{R}(L) \subset W$. If L^{-1} is unbounded on $\mathcal{R}(L)$ to V, we say the original problem $Lv = w$ is *ill-posed* or *unstable*. Such problems are not considered in this text, but there are a number of important applications (e.g. many indirect sensing devices) which fall into this category. Problems in which L^{-1} is bounded (along with L) are called *well-posed* or *stable*; they can still be ill-conditioned, as was discussed in the preceding paragraph.

2.4.3 Principle of uniform boundedness

Another widely used set of results refer to the collective boundedness of a set of linear operators.

Theorem 2.4.4 *Let $\{L_n\}$ be a sequence of bounded linear operators from a Banach space V to a normed space W. Assume for every $v \in V$, the sequence $\{L_n v\}$ is bounded. Then*

$$\sup_n \|L_n\| < \infty.$$

This theorem is often called the *principle of uniform boundedness*; see [58, p. 95] or [250, p. 172] for a proof and a more extended development. We also have the following useful variant of this principle.

Theorem 2.4.5 (Banach-Steinhaus theorem) *Let V and W be normed spaces with V being complete, and let $L, L_n \in \mathcal{L}(V, W)$. Let V_0 be a dense subspace of V. Then in order for $L_n v \to Lv \ \forall\, v \in V$, it is necessary and sufficient that*
(a) *$L_n v \to Lv \ \forall\, v \in V_0$; and*
(b) $\sup_n \|L_n\| < \infty$.

Proof. ($\Rightarrow$) Assume $L_n v \to Lv$ for all $v \in V$. Then (a) follows trivially; and (b) follows from the principle of uniform boundedness.
($\Leftarrow$) Assume (a) and (b). Denote $B = \sup_n \|L_n\|$. Let $v \in V$ and $\epsilon > 0$. By the denseness of V_0 in V, there is an element $v_\epsilon \in V_0$ such that

$$\|v - v_\epsilon\| \leq \frac{\epsilon}{3 \max\{\|L\|, B\}}.$$

Then

$$\begin{aligned} \|Lv - L_n v\| &\leq \|Lv - Lv_\epsilon\| + \|Lv_\epsilon - L_n v_\epsilon\| + \|L_n v_\epsilon - L_n v\| \\ &\leq \|L\| \, \|v - v_\epsilon\| + \|Lv_\epsilon - L_n v_\epsilon\| + \|L_n\| \, \|v_\epsilon - v\| \\ &\leq \frac{2\,\epsilon}{3} + \|Lv_\epsilon - L_n v_\epsilon\|. \end{aligned}$$

Using (a), we can find a natural number n_ϵ such that

$$\|Lv_\epsilon - L_n v_\epsilon\| \le \frac{\epsilon}{3}, \quad n \ge n_\epsilon.$$

Combining these results,

$$\|Lv - L_n v\| \le \epsilon, \quad n \ge n_\epsilon.$$

Therefore, $L_n v \to Lv$ as $n \to \infty$. □

Next, we apply Banach-Steinhaus theorem to discuss the convergence of numerical quadratures (i.e., numerical integration formulas).

2.4.4 Convergence of numerical quadratures

As an example, let us consider the convergence of numerical quadratures for the computation of the integral

$$Lv = \int_0^1 w(x)\, v(x)\, dx,$$

where w is a weighted function, $w(x) \ge 0$, $w \in L^1(0,1)$. There are several approaches to constructing numerical quadratures. One popular approach is to replace the function v by some interpolant of it, denoted here by Πv, and then define the corresponding numerical quadrature by the formula

$$\int_0^1 w(x)\, \Pi v(x)\, dx.$$

The topic of function interpolation is discussed briefly in Section 3.2. If Πv is taken to be the Lagrange polynomial interpolant of v on a uniform partition of the integration interval and the weight function $w(x) = 1$, the resulting quadratures are called *Newton-Cotes integration formulas*. It is well-known that high degree polynomial interpolation on a uniform partition leads to strong oscillations near the boundary of the interval and hence divergence of the interpolation in many cases. Correspondingly, one cannot expect the convergence of the Newton-Cotes integration formulas. To guarantee the convergence, one may use the Lagrange polynomial interpolant Πv of v on a properly chosen partition with more nodes placed near the boundary, or one may use a piecewise polynomial interpolant on a uniform partition or a partition suitably refined in areas where the integrand $w\, v$ changes rapidly. With the use of piecewise polynomial interpolants of v, we get the celebrated *trapezoidal rule* (using piecewise linear interpolation) and *Simpson's rule* (using piecewise quadratic interpolation).

A second popular approach to constructing numerical quadratures is by the *method of undetermined parameters*. We approximate the integral by a

sequence of finite sums, each of them being a linear combination of some function values (and more generally, derivative values can used in the sums as well). In other words, we let

$$Lv \approx L_n v = \sum_{i=0}^{n} w_i^{(n)} v(x_i^{(n)}) \tag{2.4.2}$$

and choose the weights $\{w_i^{(n)}\}_{i=0}^n$ and the nodes $\{x_i^{(n)}\}_{i=0}^n \subset [0,1]$ by some specific requirements. Some of the weights and nodes may be prescribed *a priori* according to the context of the applications, and the remaining ones are usually determined by requiring the quadrature be exact for polynomials of degree as high as possible. If none of the weights and nodes is prescribed, then we may choose these $2n+2$ quantities so that the quadrature is exact for any polynomial of degree less than or equal to $2n+1$. The resulting numerical quadratures are called *Gaussian quadratures*.

Detailed discussions of numerical quadratures (for the case when the weight function $w \equiv 1$) can be found in [15, Section 5.3]. Here we study the convergence of numerical quadratures in an abstract framework.

Let there be given a sequence of quadratures

$$L_n v = \sum_{i=0}^{n} w_i^{(n)} v(x_i^{(n)}), \tag{2.4.3}$$

where $0 \le x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} \le 1$ is a partition of $[0,1]$. We regard L_n as a linear functional (i.e. a continuous linear operator with scalar values, see next section) defined on $C[0,1]$ with the standard uniform norm. It is straightforward to show that

$$\|L_n\| = \sum_{i=0}^{n} |w_i^{(n)}|, \tag{2.4.4}$$

and this is left as an exercise for the reader.

As an important special case, assume the quadrature scheme L_n has a degree of precision $d(n)$, i.e. the scheme is exact for polynomials of degree less than or equal to $d(n)$,

$$L_n v = Lv \quad \forall\, v \in \mathbb{P}_{d(n)}.$$

Here $\mathbb{P}_{d(n)}$ is the space of all the polynomials of degree less than or equal to $d(n)$, and we assume $d(n) \to \infty$ as $n \to \infty$. Note that the subspace V_0 of all the polynomials is dense in $V = C[0,1]$. Then an application of the Banach-Steinhaus theorem shows that $L_n v \to Lv$ for any $v \in C[0,1]$ if and only if

$$\sup_n \sum_{i=0}^{n} |w_i^{(n)}| < \infty.$$

Continuing the discussion on the convergence of numerical quadratures, we assume all the conditions stated in the previous paragraph are valid. Additionally, we assume the weights $w_i^{(n)} \geq 0$. Then it follows that $L_n v \to Lv$ for any $v \in C[0,1]$ (Exercise 2.4.3).

From the point of view of numerical computations, it is important to have non-negative quadrature weights to avoid round-off error accumulations. It can be shown that for the Gaussian quadratures, all the quadrature weights are non-negative; and if the weight function w is positive on $(0,1)$, then the quadrature weights are positive. See [15, Section 5.3] for an extended discussion of Gaussian quadrature.

In the above discussion, we consider the situation where the degree of precision of the integration formula is increasingly large. Another situation we can consider is where the degree of precision of the integration formula does not increase and the integration formulas integrate exactly piecewise low degree polynomials corresponding to more and more integrating nodes; see Exercise 2.4.5 for such an example.

Exercise 2.4.1 Assume V is a Banach space with either of two norms $\|\cdot\|_1$ and $\|\cdot\|_2$. Suppose for some constant $c > 0$,

$$\|v\|_2 \leq c\,\|v\|_1 \quad \forall\, v \in V.$$

Show that the two norms are equivalent.

Exercise 2.4.2 Prove the formula (2.4.4).

Exercise 2.4.3 Consider the quadrature formula (2.4.3). Assume all the weights $w_i^{(n)}$ are non-negative and the quadrature formula is exact for polynomials of degree less than or equal to $d(n)$ with $d(n) \to \infty$ as $n \to \infty$. Prove the convergence of the quadrature: $L_n v \to Lv$, for all $v \in C[0,1]$.

Exercise 2.4.4 Let V and W be two Banach spaces. Assume $\{L_n\} \subset \mathcal{L}(V,W)$ is such that for any $v \in V$,

$$\lim_{n\to\infty} L_n(v) = L(v)$$

exists. Show that $L \in \mathcal{L}(V,W)$ and

$$\|L\| \leq \liminf_{n\to\infty} \|L_n\|.$$

Hint: Apply the principle of uniform boundedness (Theorem 2.4.4) on the operator sequence $\{L_n\}$.

Exercise 2.4.5 A popular family of numerical quadratures is constructed by approximating the integrand by its piecewise polynomial interpolants. We take the composite trapezoidal rule as an example. The integral to be computed is

$$Lv = \int_0^1 v(x)\,dx.$$

We divide the interval $[0,1]$ into n equal parts, and denote $x_i = i/n$, $0 \le i \le n$, as the nodes. Then we approximate v by its piecewise linear interpolant $\Pi_n v$ defined by

$$\Pi_n v(x) = n\,(x_i - x)\,v(x_{i-1}) + n\,(x - x_{i-1})\,v(x_i)$$

for $x_{i-1} \le x \le x_i$, $1 \le i \le n$. Then the composite trapezoidal rule is

$$L_n v = \int_0^1 \Pi_n v(x)\,dx = \frac{1}{n}\left[\frac{1}{2}v(x_0) + \sum_{i=1}^{n-1} v(x_i) + \frac{1}{2}v(x_n)\right].$$

Show that $L_n v \to Lv$ for any $v \in C[0,1]$.

Using piecewise polynomials of higher degrees based on non-uniform partitions of the integration interval, we can develop other useful numerical quadratures.

Exercise 2.4.6 In the formula (2.4.1), show that the inequality can be made as close as desired to equality for suitable choices of v and $\widetilde{v}$. More precisely, show that

$$\sup_{v,\widetilde{v}} \left(\frac{\|v - \widetilde{v}\|}{\|v\|} \div \frac{\|w - \widetilde{w}\|}{\|w\|} \right) = \|L\|\,\|L^{-1}\|.$$

Exercise 2.4.7 Let the numerical integration formula $L_n v$ of (2.4.2) be the Newton-Cotes formula that is described earlier in Subsection 2.4.4. It is known that there are continuous functions $v \in C[0,1]$ for which $L_n v \not\to Lv$ as $n \to \infty$. Accepting this, show that

$$\sup_n \sum_{i=0}^{n} \left| w_i^{(n)} \right| = \infty.$$

Moreover, show

$$\sum_{i=0}^{n} w_i^{(n)} = 1.$$

These results imply that the quadrature weights must be of varying sign and that they must be increasing in size as $n \to \infty$.

2.5 Linear functionals

An important special case of linear operators is when they take on scalar values. Let V be a normed space, and $W = \mathbb{K}$, the set of scalars associated with V. The elements in $\mathcal{L}(V, \mathbb{K})$ are called *linear functionals*. Since $\mathbb{K}$ is complete, $\mathcal{L}(V, \mathbb{K})$ is a Banach space. This space is usually denoted as V' and it is called the *dual space* of V. Usually we use lower case letters, such as ℓ, to denote a linear functional.

In some references, the term *linear functional* is used for the linear operators from a normed space to $\mathbb{K}$, without the functionals being necessarily bounded. In this work, since we use exclusively linear functionals which are bounded, we use the term "linear functionals" to refer to only bounded linear functionals.

Example 2.5.1 Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. It is a well-known result that for $1 \leq p < \infty$, the dual space of $L^p(\Omega)$ can be identified with $L^{p'}(\Omega)$. Here p' is the conjugate exponent of p, defined by the relation

$$\frac{1}{p} + \frac{1}{p'} = 1.$$

By convention, $p' = \infty$ when $p = 1$. In other words, given an $\ell \in (L^p(\Omega))'$, there is a function $u \in L^{p'}(\Omega)$, uniquely determined a.e., such that

$$\ell(v) = \int_\Omega u(\boldsymbol{x})v(\boldsymbol{x})\,dx \quad \forall\, v \in L^p(\Omega). \tag{2.5.1}$$

Conversely, for any $u \in L^{p'}(\Omega)$, the rule

$$v \longmapsto \int_\Omega u(\boldsymbol{x})v(\boldsymbol{x})\,dx, \quad v \in L^p(\Omega)$$

defines a bounded linear functional on $L^p(\Omega)$. It is convenient to identify $\ell \in (L^p(\Omega))'$ and $u \in L^{p'}(\Omega)$, related as in (2.5.1). Then we write

$$(L^p(\Omega))' = L^{p'}(\Omega), \quad 1 \leq p < \infty.$$

When $p = 2$, we have $p' = 2$. This special case is examined later in Example 2.5.9 from another perspective. For $p = \infty$, $p' = 1$. The dual space of $L^\infty(\Omega)$, however, is larger than the space $L^1(\Omega)$. □

All the results discussed in the previous sections for general linear operators certainly apply to linear functionals. In addition, there are useful results particular to linear functionals only.

2.5.1 An extension theorem for linear functionals

We have seen that a bounded linear operator can be extended to the closure of its domain. It is also possible to extend linear functionals defined on an arbitrary subspace to the whole space.

Theorem 2.5.2 (Hahn-Banach theorem) *Let V_0 be a subspace of a normed space V, and $\ell : V_0 \to \mathbb{K}$ be linear and bounded. Then there exists an extension $\hat{\ell} \in V'$ of ℓ with $\hat{\ell}(v) = \ell(v)\ \forall\, v \in V_0$, and $\|\hat{\ell}\| = \|\ell\|$.*

A proof can be found in [58, p. 79] or [250, p. 4]. This theorem can be proved by applying the Generalized Hahn-Banach Theorem, Theorem 2.5.5 (Exercise 2.5.1). Note that if V_0 is not dense in V, then the extension need not be unique.

Example 2.5.3 This example is important in the analysis of some numerical methods for solving integral equations. Let $V = L^\infty(0,1)$. This is the space of all cosets (or equivalence classes)

$$\boldsymbol{v} = [v] = \{w \text{ Lebesgue measurable on } [0,1] \mid w = v \text{ a.e. in } [0,1]\}$$

for which

$$\|\boldsymbol{v}\|_\infty \equiv \|v\|_\infty = \operatorname*{ess\,sup}_{0\le x\le 1} |v(x)| < \infty. \tag{2.5.2}$$

With this norm, $L^\infty(0,1)$ is a Banach space.

Let V_0 be the set of all cosets $\boldsymbol{v} = [v]$, where $v \in C[0,1]$. It is a proper subspace of $L^\infty(0,1)$. When restricted to V_0, the norm (2.5.2) is equivalent to the usual norm $\|\cdot\|_\infty$ on $C[0,1]$. It is common to write $V_0 = C[0,1]$; but this is an abuse of notation, and it is important to keep in mind the distinction between V_0 and $C[0,1]$.

Let $c \in [0,1]$, and define

$$\ell_c([v]) = v(c) \quad \forall\, v \in C[0,1]. \tag{2.5.3}$$

The linear functional $\ell_c([v])$ is well-defined on V_0. From

$$|\ell_c([v])| = |v(c)| \le \|v\|_\infty = \|\boldsymbol{v}\|_\infty, \quad \boldsymbol{v} = [v],$$

we see that $\|\ell_c\| \le 1$. By choosing $v \in C[0,1]$ with $v(c) = \|v\|_\infty$, we then obtain

$$\|\ell_c\| = 1.$$

Using Hahn-Banach Theorem, we can extend ℓ_c to $\hat{\ell}_c : L^\infty(0,1) \to \mathbb{K}$ with

$$\|\hat{\ell}_c\| = \|\ell_c\| = 1.$$

The functional $\hat{\ell}_c$ extends to $L^\infty(0,1)$ the concept of point evaluation of a function, to functions which are only Lebesgue measurable and which are not precisely defined because of being members of a coset.

Somewhat surprisingly, many desirable properties of ℓ_c are carried over to $\hat{\ell}_c$. These include the following.

- Assume $[v] \in L^\infty(0,1)$ satisfies $m \le v(x) \le M$ for almost all x in some open interval about c. Then $m \le \hat{\ell}_c([v]) \le M$.
- Assume c is a point of continuity of v. Then

$$\hat{\ell}_c([v]) = v(c),$$
$$\lim_{a\to c} \hat{\ell}_a([v]) = v(c).$$

These ideas and properties carry over to $L^\infty(D)$, with D a closed, bounded set in $\mathbb{R}^d$, $d \ge 1$, and $\boldsymbol{c} \in D$. For additional detail and the application of this extension to numerical integral equations, see [22]. □

In some applications, a stronger form of the Hahn-Banach Theorem is needed. We begin by introducing another useful concept for functionals.

Definition 2.5.4 *A functional p on a real vector space V is said to be* sublinear *if*

$$p(u+v) \le p(u) + p(v) \quad \forall\, u, v \in V,$$
$$p(\alpha v) = \alpha\, p(v) \quad \forall\, v \in V, \forall\, \alpha \ge 0.$$

We note that a semi-norm is a sublinear functional. A proof of the following result can be found in [58, p. 78] or [250, p. 2].

Theorem 2.5.5 (Generalized Hahn-Banach Theorem) *Let V be a linear space, $V_0 \subset V$ a subspace. Suppose $p : V \to \mathbb{R}$ is a sublinear functional and $\ell : V_0 \to \mathbb{R}$ a linear functional such that $\ell(v) \le p(v)$ for all $v \in V_0$. Then ℓ can be extended to V such that $\ell(v) \le p(v)$ for all $v \in V$.*

Note that $p(v) = c\,\|v\|_V$, c a positive constant, is a sublinear functional on V. With this choice of p, we obtain the original Hahn-Banach Theorem. Another useful consequence of the generalized Hahn-Banach Theorem is the following result, its proof being left as Exercise 2.5.2.

Corollary 2.5.6 *Let V be a normed space. For any $0 \ne v \in V$, there exists $\ell_v \in V'$ such that $\|\ell_v\| = 1$ and $\ell_v(v) = \|v\|$.*

We can use Corollary 2.5.6 to characterize the norm in a normed linear space.

Corollary 2.5.7 *Let V be a normed space. Then for any $v \in V$,*

$$\|v\| = \sup\{|\ell(v)| \mid \ell \in V',\ \|\ell\| = 1\}. \tag{2.5.4}$$

Proof. The result is obvious for $v = 0$. Assume $v \ne 0$. For any $\ell \in V'$ with $\|\ell\| = 1$, we have

$$|\ell(v)| \le \|\ell\|\,\|v\| = \|v\|.$$

By Corollary 2.5.6, we have an $\ell_v \in V'$ with $\|\ell_v\| = 1$ such that $\ell_v(v) = \|v\|$. Hence, the equality (2.5.4) holds. □

2.5.2 The Riesz representation theorem

On Hilbert spaces, linear functionals are limited in the forms they can take. The following theorem makes this more precise; and the result is one used in developing the solvability theory for some important partial differential equations and boundary integral equations. The theorem also provides a tool for introducing the concept of the adjoint of a linear operator in the next section.

Theorem 2.5.8 (Riesz representation theorem) *Let V be a real or complex Hilbert space, $\ell \in V'$. Then there is a unique $u \in V$ for which*

$$\ell(v) = (v, u) \quad \forall\, v \in V. \tag{2.5.5}$$

In addition,

$$\|\ell\| = \|u\|. \tag{2.5.6}$$

Proof. Assuming the existence of u, we first prove its uniqueness. Suppose $\tilde{u} \in V$ satisfies

$$\ell(v) = (v, u) = (v, \tilde{u}) \quad \forall\, v \in V.$$

Then

$$(v, u - \tilde{u}) = 0 \quad \forall\, v \in V.$$

Take $v = u - \tilde{u}$. Then $\|u - \tilde{u}\| = 0$, which implies $u = \tilde{u}$.

We give two derivations of the existence of u, both for the case of a real Hilbert space.

Standard proof of existence. Denote

$$N = \mathcal{N}(\ell) = \{v \in V \mid \ell(v) = 0\},$$

which is a subspace of V. If $N = V$, then $\|\ell\| = 0$, and we may take $u = 0$.

Now suppose $N \neq V$. Then there exists at least one $v_0 \in V$ such that $\ell(v_0) \neq 0$. It is possible to decompose V as the direct sum of N and $N^{\perp}$ (see Section 3.6). From this, we have the decomposition $v_0 = v_1 + v_2$ with $v_1 \in N$ and $v_2 \in N^{\perp}$. Then $\ell(v_2) = \ell(v_0) \neq 0$.

For any $v \in V$, we have the property

$$\ell\left(v - \frac{\ell(v)}{\ell(v_2)} v_2\right) = 0.$$

Thus

$$v - \frac{\ell(v)}{\ell(v_2)} v_2 \in N,$$

and in particular, it is orthogonal to v_2:

$$\left(v - \frac{\ell(v)}{\ell(v_2)} v_2, v_2\right) = 0,$$

i.e.,

$$\ell(v) = \left(v, \frac{\ell(v_2)}{\|v_2\|^2} v_2\right).$$

In other words, we may take u to be $[\ell(v_2)/\|v_2\|^2]v_2$.

Proof using a minimization principle. From Theorem 3.3.12 in Chapter 3, we know the problem

$$\inf_{v \in V} \left[\frac{1}{2}\|v\|^2 - \ell(v)\right]$$

has a unique solution $u \in V$. The solution u is characterized by the relation (2.5.5).

We complete the proof of the theorem by showing (2.5.6). From (2.5.5) and the Schwarz inequality,

$$|\ell(v)| \le \|u\|\,\|v\| \quad \forall\, v \in V.$$

Hence

$$\|\ell\| \le \|u\|.$$

Let $v = u$ in (2.5.5). Then

$$\ell(u) = \|u\|^2$$

and

$$\|\ell\| = \sup_{v \neq 0} \frac{|\ell(v)|}{\|v\|} \ge \frac{|\ell(u)|}{\|u\|} \ge \|u\|.$$

Therefore, (2.5.6) holds. □

This theorem seems very straightforward, and its proof seems fairly simple. Nonetheless, this is a fundamental tool in the solvability theory for elliptic partial differential equations, as we see later in Chapter 8.

Example 2.5.9 Let $\Omega \subset \mathbb{R}^d$ be open bounded. $V = L^2(\Omega)$ is a Hilbert space. By the Riesz representation theorem, there is a one-to-one correspondence between V' and V by the relation (2.5.5). We can identify $\ell \in V'$ with $u \in V$ related by (2.5.5). In this sense, $(L^2(\Omega))' = L^2(\Omega)$. □

For the space $L^2(\Omega)$, the element u in (2.5.5) is almost always immediately apparent. But for spaces such as the Sobolev space $H^1(a,b)$ introduced in Example 1.3.7 of Chapter 1, the determination of u of (2.5.5) is often not as obvious. As an example, define $\ell \in (H^1(a,b))'$ by

$$\ell(v) = v(c), \quad v \in H^1(a,b) \tag{2.5.7}$$

for some $c \in [a,b]$. This linear functional can be shown to be well-defined (Exercise 2.5.5). From the Riesz representation theorem, there is a unique $u \in H^1(a,b)$ such that

$$\int_a^b [u'(x)\,v'(x) + u(x)\,v(x)]\,dx = v(c) \quad \forall\, v \in H^1(a,b). \tag{2.5.8}$$

The element u is the generalized solution of the boundary value problem

$$\begin{aligned} -u'' + u &= \delta(x-c) \ \text{ in } (a,b), \\ u'(a) &= u'(b) = 0, \end{aligned}$$

where $\delta(x-c)$ is the Dirac δ-function at c. This boundary value problem can be written equivalently as

$$\begin{aligned} -u'' + u &= 0 \ \text{ in } (a,c)\cup(c,b), \\ u'(a) &= u'(b) = 0, \\ u(c-) &= u(c+), \\ u'(c-) - u'(c+) &= 1. \end{aligned}$$

Exercise 2.5.1 Use Theorem 2.5.5 to prove Theorem 2.5.2.
Hint: Take $p(v) = \|\ell\|\,\|v\|$.

Exercise 2.5.2 Prove Corollary 2.5.6.

Exercise 2.5.3 Let us discuss the Riesz representation theorem for a finite-dimensional inner product space. Suppose ℓ is a linear functional on $\mathbb{R}^d$. Show directly the formula

$$\ell(\boldsymbol{x}) = (\boldsymbol{x}, \boldsymbol{l}) \quad \forall\, \boldsymbol{x} \in \mathbb{R}^d.$$

Determine the vector $\boldsymbol{l} \in \mathbb{R}^d$.

Exercise 2.5.4 Suppose $\{e_j\}_{j=1}^{\infty}$ is an orthonormal basis of a real Hilbert space V. Show that the element $u \in V$ satisfying (2.5.5) is given by the formula

$$u = \sum_{j=1}^{\infty} \ell(e_j)\, e_j.$$

Exercise 2.5.5 Show that the functional defined in (2.5.7) is linear and bounded on $H^1(a,b)$.
Hint: Use the following results. For any $f \in H^1(a,b)$, f is continuous on $[a,b]$, and therefore,

$$\int_a^b f(x)\,dx = f(\zeta)$$

for some $\zeta \in (a,b)$. In addition,

$$f(c) = f(\zeta) + \int_{\zeta}^{c} f'(x)\,dx.$$

Exercise 2.5.6 Find the function u of (2.5.8) through the boundary value problem it satisfies.

2.6 Adjoint operators

The notion of an adjoint operator is a generalization of the matrix transpose to infinite dimensional spaces. First let us derive a defining property for the

matrix transpose. Let $A \in \mathbb{R}^{m\times n}$, which is viewed as a linear continuous operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. We use the conventional Euclidean inner products for the spaces $\mathbb{R}^n$ and $\mathbb{R}^m$. Then

$$\boldsymbol{y}^T A\boldsymbol{x} = (A\boldsymbol{x}, \boldsymbol{y})_{\mathbb{R}^m}, \quad \boldsymbol{x}^T A^T \boldsymbol{y} = (\boldsymbol{x}, A^T\boldsymbol{y})_{\mathbb{R}^n} \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{y} \in \mathbb{R}^m.$$

Since $\boldsymbol{y}^T A\boldsymbol{x}$ is a real number, $\boldsymbol{y}^T A\boldsymbol{x} = (\boldsymbol{y}^T A\boldsymbol{x})^T = \boldsymbol{x}^T A^T\boldsymbol{y}$. We observe that the transpose (or adjoint) A^T is uniquely defined by the property

$$(A\boldsymbol{x}, \boldsymbol{y})_{\mathbb{R}^m} = (\boldsymbol{x}, A^T\boldsymbol{y})_{\mathbb{R}^n} \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{y} \in \mathbb{R}^m.$$

Turning now to the general situation, assume V and W are Hilbert spaces, $L \in \mathcal{L}(V, W)$. Let us use the Riesz representation theorem to define a new operator $L^* : W \to V$, called the *adjoint* of L. For simplicity, we assume in this section that $\mathbb{K} = \mathbb{R}$ for the set of scalars associated with W and V. Given $w \in W$, define a linear functional $\ell_w \in V'$ by

$$\ell_w(v) = (Lv, w)_W \quad \forall\, v \in V.$$

This linear functional is bounded because

$$|\ell_w(v)| \le \|Lv\|\,\|w\| \le \|L\|\,\|v\|\,\|w\|$$

and so

$$\|\ell_w\| \le \|L\|\,\|w\|.$$

By the Riesz representation theorem, there is a uniquely determined element, denoted by $L^*(w) \in V$ such that

$$\ell_w(v) = (v, L^*(w)) \quad \forall\, v \in V.$$

We write

$$(Lv, w)_W = (v, L^*(w))_V \quad \forall\, v \in V,\ w \in W.$$

We first show that L^* is linear. Let $w_1, w_2 \in W$, and consider the linear functionals

$$\begin{aligned} \ell_1(v) &= (Lv, w_1)_W = (v, L^*(w_1))_V, \\ \ell_2(v) &= (Lv, w_2)_W = (v, L^*(w_2))_V \end{aligned}$$

for any $v \in V$. Add these relations,

$$(Lv, w_1 + w_2)_W = (v, L^*(w_1) + L^*(w_2))_V \quad \forall\, v \in V.$$

By definition,

$$(Lv, w_1 + w_2)_W = (v, L^*(w_1 + w_2))_V;$$

so

$$(v, L^*(w_1 + w_2))_V = (v, L^*(w_1) + L^*(w_2))_V \quad \forall\, v \in V.$$

This implies

$$L^*(w_1 + w_2) = L^*(w_1) + L^*(w_2).$$

By a similar argument, for any $\alpha \in \mathbb{R}$, any $w \in W$,

$$L^*(\alpha\, w) = \alpha\, L^*(w).$$

Hence L^* is linear and we write $L^*(w) = L^* w$, and the defining relation is

$$(Lv, w)_W = (v, L^* w)_V \quad \forall\, v \in V,\, w \in W. \tag{2.6.1}$$

Then we show the boundedness of L^*. We have

$$\|L^* w\| = \|\ell_w\| \le \|L\|\, \|w\| \quad \forall\, w \in W.$$

Thus

$$\|L^*\| \le \|L\| \tag{2.6.2}$$

and L^* is bounded. Let us show that actually the inequality in (2.6.2) can be replaced by an equality. For this, we consider the adjoint of L^*, defined by the relation

$$(L^* w, v)_V = (w, (L^*)^* v)_W \quad \forall\, v \in V,\, w \in W.$$

Thus

$$(w, (L^*)^* v)_W = (w, Lv)_W \quad \forall\, v \in V,\, w \in W.$$

By writing this as $(w, (L^*)^* v - Lv)_W = 0$ and letting $w = (L^*)^* v - Lv$, we obtain

$$(L^*)^* v = Lv \quad \forall\, v \in V.$$

Hence

$$(L^*)^* = L. \tag{2.6.3}$$

We then apply (2.6.1) to L^* to obtain

$$\|L\| = \|(L^*)^*\| \le \|L^*\|.$$

Combining this with (2.6.2), we have

$$\|L^*\| = \|L\|. \tag{2.6.4}$$

From the above derivation, we see that for a continuous linear operator between Hilbert spaces, the adjoint of its adjoint is the operator itself.

In the special situation $V = W$ and $L = L^*$, we say L is a *self-adjoint* operator. When L is a self-adjoint operator from $\mathbb{R}^n$ to $\mathbb{R}^n$, it is represented by a symmetric matrix in $\mathbb{R}^{n\times n}$. Equations of the form $Lv = w$ with L self-adjoint occur in many important physical settings, and the study of them forms a large and important area within functional analysis.

Example 2.6.1 Let $V = W = L^2(a,b)$ with the real numbers as scalars and the standard norm $\|\cdot\|_2$. Consider the linear integral operator

$$Kv(x) = \int_a^b k(x,y)\, v(y)\, dy, \quad a \le x \le b,$$

where the kernel function satisfies the condition

$$B \equiv \left[\int_a^b \int_a^b |k(x,y)|^2 dx\, dy\right]^{1/2} < \infty.$$

For any $v \in L^2(a,b)$,

$$\begin{aligned}\|Kv\|_2^2 &= \int_a^b \left|\int_a^b k(x,y)\, v(y)\, dy\right|^2 dx \\ &\le \int_a^b \left[\int_a^b |k(x,y)|^2 dy\right]\left[\int_a^b |v(y)|^2 dy\right] dx \\ &= B^2 \|v\|_2^2.\end{aligned}$$

Thus,

$$\|Kv\|_2 \le B\, \|v\|_2 \quad \forall\, v \in L^2(a,b),$$

and then

$$\|K\| \le B.$$

Hence we see that K is a continuous linear operator on $L^2(a,b)$.

Now let us find the adjoint of K. By the defining relation (2.6.1),

$$\begin{aligned}(v, K^*w) &= (Kv, w) \\ &= \int_a^b w(x) \left[\int_a^b k(x,y)\, v(y)\, dy\right] dx \\ &= \int_a^b \left[\int_a^b k(x,y)\, w(x)\, dx\right] v(y)\, dy\end{aligned}$$

for any $v, w \in L^2(a,b)$. This implies

$$K^*v(y) = \int_a^b k(x,y)\, v(x)\, dx \quad \forall\, v \in L^2(a,b).$$

The integral operator K is self-adjoint if and only if $k(x,y) = k(y,x)$. □

Given a Hilbert space V, the set of self-adjoint operators on V form a subspace of $\mathcal{L}(V)$. Indeed the following result is easy to verify.

Proposition 2.6.2 *If $L_1, L_2 \in \mathcal{L}(V)$ are self-adjoint, then for any real scalars α_1 and α_2, the operator $\alpha_1 L_1 + \alpha_2 L_2$ is self-adjoint.*

Proof. From Exercise 2.6.1, we have

$$(\alpha_1 L_1 + \alpha_2 L_2)^* = \alpha_1 L_1^* + \alpha_2 L_2^*.$$

Since L_1 and L_2 are self-adjoint,

$$(\alpha_1 L_1 + \alpha_2 L_2)^* = \alpha_1 L_1 + \alpha_2 L_2.$$

Hence $\alpha_1 L_1 + \alpha_2 L_2$ is self-adjoint. □

Proposition 2.6.3 *Assume $L_1, L_2 \in \mathcal{L}(V)$ are self-adjoint. Then $L_1 L_2$ is self-adjoint if and only if $L_1 L_2 = L_2 L_1$.*

Proof. Since L_1 and L_2 are self-adjoint, we have

$$(L_1 L_2 u, v) = (L_2 u, L_1 v) = (u, L_2 L_1 v) \quad \forall\, u, v \in V.$$

Thus

$$(L_1 L_2)^* = L_2 L_1.$$

It follows that $L_1 L_2$ is self-adjoint if and only if $L_1 L_2 = L_2 L_1$ is valid. □

Corollary 2.6.4 *Suppose $L \in \mathcal{L}(V)$ is self-adjoint. Then for any non-negative integer n, L^n is self-adjoint (by convention, $L^0 = I$, the identity operator). Consequently, for any polynomial $p(x)$ with real coefficients, the operator $p(L)$ is self-adjoint.*

We have a useful characterization of the norm of a self-adjoint operator.

Theorem 2.6.5 *Let $L \in \mathcal{L}(V)$ be self-adjoint. Then*

$$\|L\| = \sup_{\|v\|=1} |(Lv, v)|. \tag{2.6.5}$$

Proof. Denote $M = \sup_{\|v\|=1} |(Lv, v)|$. First for any $v \in V$, $\|v\| = 1$, we have

$$|(Lv, v)| \le \|Lv\|\, \|v\| \le \|L\|.$$

So

$$M \le \|L\|. \tag{2.6.6}$$

Now for any $u, v \in V$, we have the identity

$$(Lu, v) = \frac{1}{4}\left[(L(u+v), u+v) - (L(u-v), u-v)\right].$$

Thus

$$|(Lu, v)| \leq \frac{M}{4}\left(\|u+v\|^2 + \|u-v\|^2\right) = \frac{M}{2}\left(\|u\|^2 + \|v\|^2\right).$$

For $u \in V$ with $Lu \neq 0$, we take $v = (\|u\|/\|Lu\|)\, Lu$ in the above inequality to obtain

$$\|u\|\,\|Lu\| \leq M\,\|u\|^2,$$

i.e.,

$$\|Lu\| \leq M\,\|u\|.$$

Obviously, this inequality also holds if $Lu = 0$. Hence,

$$\|Lu\| \leq M\,\|u\| \quad \forall\, u \in V,$$

and we see that $\|L\| \leq M$. This inequality and (2.6.6) imply (2.6.5). □

Exercise 2.6.1 Prove the following properties for adjoint operators.

$$\begin{aligned}(\alpha_1 L_1 + \alpha_2 L_2)^* &= \alpha_1 L_1^* + \alpha_2 L_2^*, \quad \alpha_1, \alpha_2 \text{ real},\\ (L_1 L_2)^* &= L_2^* L_1^*,\\ (L^*)^* &= L.\end{aligned}$$

Exercise 2.6.2 Define $K : L^2(0,1) \to L^2(0,1)$ by

$$Kf(x) = \int_0^x k(x,y) f(y)\, dy, \quad 0 \leq x \leq 1,\ f \in L^2(0,1),$$

with $k(x,y)$ continuous for $0 \leq y \leq x \leq 1$. Show K is a bounded operator. What is K^*? To what extent can the assumption of continuity of $k(x,y)$ be made less restrictive?

Exercise 2.6.3 The right-hand side of (2.6.5) defines a quantity $|||L|||$ for a general linear continuous operator $L \in \mathcal{L}(V)$. Prove the inequality

$$|(Lu, v)| \leq |||L|||\,\|u\|\,\|v\|.$$

Hint: First consider the case $\|u\| = \|v\| = 1$.

2.7 Weak convergence and weak compactness

We recall from Definition 1.2.8 that in a normed space V, a sequence $\{u_n\}_n$ is said to converge to an element $u \in V$ if

$$\lim_{n\to\infty} \|u_n - u\| = 0.$$

Such convergence is also called strong convergence or convergence in norm. We write $u_n \to u$ in V as $n \to \infty$ to express this convergence property.

In this section, we introduce convergence in a weak sense.

Definition 2.7.1 *Let V be a normed space, V' its dual space. A sequence $\{u_n\} \subset V$* converges weakly *to $u \in V$ if*

$$\ell(u_n) \to \ell(u) \quad \text{as } n \to \infty, \ \forall\, \ell \in V'.$$

In this case, we write $u_n \rightharpoonup u$ as $n \to \infty$.

From the definition, it is easy to see that strong convergence implies weak convergence.

Proposition 2.7.2 *Let V be a normed space, and assume $\{u_n\}$ converges weakly to u in V. Then*

$$\sup_n \|u_n\| < \infty.$$

Proof. From the sequence $\{u_n\} \subset V$, we define a sequence $\{u_n^*\} \subset (V')'$ by

$$u_n^*(\ell) = \ell(u_n) \quad \forall\, \ell \in V'.$$

Then $\{u_n^*\}$ is a sequence of bounded linear operators defined on the Banach space V', and for every $\ell \in V'$, the sequence $\{u_n^*(\ell)\}$ is bounded since $u_n^*(\ell) = \ell(u_n)$ converges. By the principle of uniform boundedness (Theorem 2.4.4), we know that

$$\sup_n \|u_n^*\| < \infty.$$

Apply Corollary 2.5.7, for any n,

$$\begin{aligned} \|u_n\| &= \sup\{|\ell(u_n)| \mid \ell \in V',\ \|\ell\| = 1\} \\ &= \sup\{|u_n^*(\ell)| \mid \ell \in V',\ \|\ell\| = 1\} \\ &\le \|u_n^*\|. \end{aligned}$$

Hence, $\sup_n \|u_n\| \le \sup_n \|u_n^*\| < \infty$. □

Thus, a weakly convergent sequence must be bounded.

Example 2.7.3 Let $f \in L^2(0, 2\pi)$. Then we know from Parseval's equality (1.3.10) that the Fourier series of f converges in $L^2(0, 2\pi)$. Therefore the Fourier coefficients converge to zero, and in particular,

$$\int_0^{2\pi} f(x)\, \sin(nx)\, dx \to 0 \quad \forall\, f \in L^2(0, 2\,\pi).$$

This result is known as the *Riemann-Lebesgue Lemma*. Thus the sequence $\{\sin(nx)\}_{n\ge 1}$ converges weakly to 0 in $L^2(0, 2\,\pi)$. But certainly the sequence $\{\sin(nx)\}_{n\ge 1}$ does not converge strongly to 0 in $L^2(0, 2\,\pi)$. □

Strong convergence implies weak convergence, but not vice versa as Example 2.7.3 shows, unless the space V is finite-dimensional. In a finite-dimensional space, it is well-known that a bounded sequence has a convergent subsequence (see Theorem 1.6.2). In an infinite-dimensional space, we expect only a weaker property; but even the weaker property is still useful in proving many existence results.

In the proof of Proposition 2.7.2, we have used the bidual $V'' = (V')'$. The bidual V'' of a normed space V is the dual of its dual V' with the corresponding dual norm. The mapping $J : V \to V''$ through the relation

$$\langle Jv, \ell \rangle_{V'' \times V'} = \langle \ell, v \rangle_{V' \times V} \quad \forall\, v \in V,\ \forall\, \ell \in V'$$

defines an isometric isomorphism between V and $J(V) \subset V''$. The isometry refers to the equality

$$\|Jv\|_{V''} = \|v\|_V \quad \forall\, v \in V.$$

This equality is proved as follows: Easily, $\|Jv\|_{V''} \leq \|v\|_V$, and by an application of Corollary 2.5.6, $\|Jv\|_{V''} \geq \|v\|_V$. We identify $J(V)$ with V.

Definition 2.7.4 *A normed space V is said to be* reflexive *if $(V')' = V$.*

An immediate consequence of this definition is that a reflexive normed space must be complete (i.e. a Banach space). By the Riesz representation theorem, it is relatively straightforward to show that any Hilbert space is reflexive.

The most important property of a reflexive Banach space is its weak compactness, given in the next theorem. It is fundamental in the development of an existence theory for abstract optimization problems (see Section 3.3). A proof is given in [58, p. 132] and [250, p. 64].

Theorem 2.7.5 *A Banach space V is reflexive if and only if any bounded sequence in V has a subsequence weakly converging to an element in V.*

Let $\Omega \subset \mathbb{R}^d$ be open and bounded. Recall from Example 2.5.1 that for $p \in (1, \infty)$, the dual space of $L^p(\Omega)$ is $L^{p'}(\Omega)$, where p' is the conjugate of p defined by the relation $1/p' + 1/p = 1$. Therefore,

$$(L^p(\Omega)')' = \left(L^{p'}(\Omega)\right)' = L^p(\Omega),$$

i.e., if $p \in (1, \infty)$, then the space $L^p(\Omega)$ is reflexive. Consequently, the above theorem implies the following: If $\{u_n\}$ is a bounded sequence in $L^p(\Omega)$, i.e. $\sup_n \|u_n\|_{L^p(\Omega)} < \infty$, then we can find a subsequence $\{u_{n'}\} \subset \{u_n\}$ and a function $u \in L^p(\Omega)$ such that

$$\lim_{n' \to \infty} \int_\Omega u_{n'}(\boldsymbol{x})\, v(\boldsymbol{x})\, dx = \int_\Omega u(\boldsymbol{x})\, v(\boldsymbol{x})\, dx \quad \forall\, v \in L^{p'}(\Omega).$$

The space $L^1(\Omega)$ is not reflexive since $(L^1(\Omega))' = L^\infty(\Omega)$, but $(L^\infty(\Omega))'$ strictly contains $L^1(\Omega)$ (Example 2.5.1). To state a result on weak compactness of the space $L^1(\Omega)$, we need the concept of uniform integrability. We say a set $U \subset L^1(\Omega)$ is *uniformly integrable* if U is bounded and for any $\varepsilon > 0$, there is a $\delta = \delta(\varepsilon) > 0$ such that for any measurable subset $\Omega_1 \subset \Omega$ with measure $|\Omega_1| < \delta$, we have

$$\sup_{v \in U} \int_{\Omega_1} |v(\boldsymbol{x})|\, dx < \varepsilon.$$

Then from Dunford-Pettis theorem ([26, Subsection 2.4.5]), U is uniformly integrable if and only if each sequence $\{u_n\}_n \subset U$ has a subsequence $\{u_{n'}\}_{n'}$ converging weakly to some element $u \in L^1(\Omega)$:

$$\lim_{n' \to \infty} \int_\Omega u_{n'}(\boldsymbol{x})\, v(\boldsymbol{x})\, dx = \int_\Omega u(\boldsymbol{x})\, v(\boldsymbol{x})\, dx \quad \forall\, v \in L^\infty(\Omega).$$

A necessary and sufficient condition for $U \subset L^1(\Omega)$ to be uniformly integrable is that there exists a function $f : [0, \infty) \to [0, \infty)$ such that

$$\frac{f(s)}{s} \to \infty \text{ as } s \to \infty, \quad \text{and} \quad \sup_{v \in U} \inf_{\Omega} f(|v(\boldsymbol{x})|)\, dx < \infty.$$

Finally, we introduce the concepts of strong convergence and *weak-** convergence of a sequence of linear operators.

Definition 2.7.6 *Let V and W be normed spaces. A sequence of linear operators $\{L_n\}$ from V to W is said to* converge strongly *to a linear operator $L : V \to W$ if*

$$\lim_{n \to \infty} \|L - L_n\| = 0.$$

In this case, we write $L_n \to L$ as $n \to \infty$. We say $\{L_n\}$ converges weak-* *to L and write $L_n \rightharpoonup^* L$ if*

$$\lim_{n \to \infty} L_n v = Lv \quad \forall\, v \in V.$$

We also say $\{L_n\}$ converges pointwise to L on V.

For a reflexive Banach space V, a sequence $\{L_n\} \subset V'$ converges weak-* to L if and only if the sequence converges weakly to L in V'.

We end this section with the following equivalent definitions of weak convergence in $L^p(\Omega)$ for $1 \le p < \infty$ and weak-* convergence in $L^\infty(\Omega)$: a sequence $\{u_n\}_n \subset L^p(\Omega)$ weakly converges to $u \in L^p(\Omega)$ if

$$\lim_{n \to \infty} \int_\Omega u_n(\boldsymbol{x})\, v(\boldsymbol{x})\, dx = \int_\Omega u(\boldsymbol{x})\, v(\boldsymbol{x})\, dx \quad \forall\, v \in L^{p'}(\Omega);$$

a sequence $\{u_n\}_n \subset L^\infty(\Omega)$ weakly-* converges to $u \in L^\infty(\Omega)$ if

$$\lim_{n\to\infty} \int_\Omega u_n(\boldsymbol{x})\, v(\boldsymbol{x})\, dx = \int_\Omega u(\boldsymbol{x})\, v(\boldsymbol{x})\, dx \quad \forall\, v \in L^1(\Omega).$$

Exercise 2.7.1 Consider the linear operators from $C[a,b]$ to $\mathbb{R}$ defined by

$$Lv = \int_a^b v(x)\, dx$$

and

$$L_n v = \frac{b-a}{n} \sum_{i=1}^{n} v\left(a + \frac{b-a}{n}\, i\right), \quad n = 1, 2, \ldots.$$

We recognize that $\{L_n v\}$ is a sequence of Riemann sums for the integral Lv. Show that $L_n \rightharpoonup^* L$ but $L_n \not\to L$.

Exercise 2.7.2 If $u_n \rightharpoonup u$, then

$$\|u\| \le \liminf_{n\to\infty} \|u_n\|.$$

Show by an example that in general,

$$\|u\| \ne \liminf_{n\to\infty} \|u_n\|.$$

Exercise 2.7.3 Show that in an inner product space,

$$u_n \to u \quad \Longleftrightarrow \quad u_n \rightharpoonup u \text{ and } \|u_n\| \to \|u\|.$$

Exercise 2.7.4 The equivalence stated in Exercise 2.7.3 can be extended to a class of Banach spaces known as uniformly convex Banach spaces. A Banach space V is said to be uniformly convex if for any sequence $\{u_n\} \subset V$ and $\{v_n\} \subset V$ such that $\|u_n\| = \|v_n\| = 1$ for any n, the following implication holds as $n \to \infty$:

$$\|u_n + v_n\| \to 2 \quad \Longrightarrow \quad \|u_n - v_n\| \to 0.$$

(a) Show that a Hilbert space is uniformly convex.
(b) Apply the Clarkson inequalities (1.5.5) and (1.5.6) to show that $L^p(\Omega)$, $1 < p < \infty$, is uniformly convex.
(c) Show that in a uniformly convex Banach space,

$$u_n \to u \quad \Longleftrightarrow \quad u_n \rightharpoonup u \text{ and } \|u_n\| \to \|u\|.$$

Hint: For the "$\Leftarrow$" part, consider $u_n/\|u_n\|$ and $u/\|u\|$, and note that $u_n/\|u_n\| \rightharpoonup u/\|u\|$.

Exercise 2.7.5 Recall the following Riemann-Lebesgue Theorem ([61, Theorem 2.1.5]): Let $1 \le p \le \infty$ and $\Omega = (a_1, b_1) \times \cdots \times (a_d, b_d)$ for real numbers $a_i < b_i$, $1 \le i \le d$. For any $u \in L^p(\Omega)$, extend it by periodicity from Ω to $\mathbb{R}^d$ and define a sequence of functions $\{u_n\}_{n\ge 1}$ by $u_n(\boldsymbol{x}) = u(n\boldsymbol{x})$. Denote

$\overline{u} = \int_\Omega u(\boldsymbol{x})\,dx/\text{meas}(\Omega)$ for the mean value of u over Ω. Then $u_n \rightharpoonup \overline{u}$ in $L^p(\Omega)$ if $1 \le p < \infty$, and $u_n \rightharpoonup^* \overline{u}$ in $L^\infty(\Omega)$ if $p = \infty$.

Apply this result to determine the weak limits of the following sequences: $\{\sin nx\}_{n\ge1} \subset L^p(0,\pi)$; $\{\sin nx\}_{n\ge1} \subset L^p(0,\pi/2)$; $\{\sin^2 nx\}_{n\ge1} \subset L^p(0,2\pi)$; $\{u_n(x)\}_{n\ge1} \subset L^p(0,1)$ with $u_n(x) = u(nx)$ and u is a piecewise constant function: $u(x) = c_i$ for $x \in ((i-1)/m, i/m)$, $1 \le i \le m$.

2.8 Compact linear operators

When V is a finite dimensional linear space and $A : V \to V$ is linear, the equation $Au = w$ has a well-developed solvability theory. To extend these results to infinite dimensional spaces, we introduce the concept of a *compact operator* K and then we present a theory for operator equations $Au = w$ in which $A = I - K$. Equations of the form

$$u - Ku = f \tag{2.8.1}$$

are called "equations of the second kind", and generally K is assumed to have special properties. The main idea is that compact operators are in some sense closely related to finite-dimensional operators, i.e. operators with a finite-dimensional range. If K is truly finite-dimensional, in a sense we define below, then (2.8.1) can be reformulated as a finite system of linear equations and solved exactly. If K is compact, then it is close to being finite-dimensional; and the solvability theory of (2.8.1) is similar to that for the finite-dimensional case.

In the following, recall the discussion in Section 1.6.

Definition 2.8.1 *Let V and W be normed spaces, and let $K : V \to W$ be linear. Then K is* compact *if the set $\{Kv \mid \|v\|_V \le 1\}$ has compact closure in W. This is equivalent to saying that for every bounded sequence $\{v_n\} \subset V$, the sequence $\{Kv_n\}$ has a subsequence that is convergent to some point in W. Compact operators are also called* completely continuous operators.

There are other definitions for a compact operator, but the above is the one used most commonly. In the definition, the spaces V and W need not be complete; but in virtually all applications, they are complete. With completeness, some of the proofs of the properties of compact operators become simpler, and we will always assume V and W are complete (i.e. Banach spaces) when dealing with compact operators.

2.8.1 Compact integral operators on $C(D)$

Let D be a closed bounded set in $\mathbb{R}^d$. The space $C(D)$ is to have the norm $\|\cdot\|_\infty$. Given a function $k : D \times D \to \mathbb{R}$, we define

$$Kv(\boldsymbol{x}) = \int_D k(\boldsymbol{x}, \boldsymbol{y}) v(\boldsymbol{y})\, dy, \quad \boldsymbol{x} \in D,\ v \in C(D). \tag{2.8.2}$$

We want to formulate conditions under which $K : C(D) \to C(D)$ is both bounded and compact. We assume $k(\boldsymbol{x}, \boldsymbol{y})$ is integrable as a function of $\boldsymbol{y}$, for all $\boldsymbol{x} \in D$, and further we assume the following.
(A_1) $\lim_{h \to 0} \omega(h) = 0$, with

$$\omega(h) \equiv \sup_{\substack{\boldsymbol{x}, \boldsymbol{z} \in D \\ \|\boldsymbol{x} - \boldsymbol{z}\| \le h}} \int_D |k(\boldsymbol{x}, \boldsymbol{y}) - k(\boldsymbol{z}, \boldsymbol{y})|\, dy. \tag{2.8.3}$$

Here, $\|\boldsymbol{x} - \boldsymbol{z}\|$ denotes the Euclidean length of $\boldsymbol{x} - \boldsymbol{z}$.

(A_2)

$$\sup_{\boldsymbol{x} \in D} \int_D |k(\boldsymbol{x}, \boldsymbol{y})|\, dy < \infty. \tag{2.8.4}$$

By (A_1), if $v(\boldsymbol{y})$ is bounded and integrable, then $Kv(\boldsymbol{x})$ is continuous and

$$|Kv(\boldsymbol{x}) - Kv(\boldsymbol{y})| \le \omega(\|\boldsymbol{x} - \boldsymbol{y}\|)\, \|v\|_\infty. \tag{2.8.5}$$

Using (A_2), we have boundedness of K, with its norm

$$\|K\| = \max_{\boldsymbol{x} \in D} \int_D |k(\boldsymbol{x}, \boldsymbol{y})|\, dy. \tag{2.8.6}$$

To discuss compactness of K, we first need to identify the compact sets in $C(D)$. To do this, we apply Arzela-Ascoli theorem (Theorem 1.6.3). Consider the set $S = \{Kv \mid v \in C(D),\ \|v\|_\infty \le 1\}$. This set is uniformly bounded, since $\|Kv\|_\infty \le \|K\|\, \|v\|_\infty \le \|K\|$ for any $v \in S$. In addition, S is equicontinuous from (2.8.5). Thus S has compact closure in $C(D)$, and K is a compact operator on $C(D)$ to $C(D)$.

What kernel functions k satisfy (A_1) and (A_2)? Easily, these assumptions are satisfied if $k(\boldsymbol{x}, \boldsymbol{y})$ is a continuous function of $(\boldsymbol{x}, \boldsymbol{y}) \in D$. In addition, let $D = [a, b]$ and consider

$$Kv(x) = \int_a^b \log|x - y|\, v(y)\, dy \tag{2.8.7}$$

and

$$Kv(x) = \int_a^b \frac{1}{|x - y|^\beta} v(y)\, dy \tag{2.8.8}$$

with $\beta < 1$. These operators K can be shown to satisfy (A_1)–(A_2), although we omit the proof. Later we show by other means that these are compact operators. An important and related example is

$$Kv(\boldsymbol{x}) = \int_D \frac{1}{\|\boldsymbol{x} - \boldsymbol{y}\|^\beta} v(\boldsymbol{y})\, dy, \quad \boldsymbol{x} \in D,\ v \in C(D).$$

The set $D \subset \mathbb{R}^d$ is assumed to be closed, bounded, and have a non-empty interior. This operator satisfies (A_1)–(A_2) provided $\beta < d$, and therefore K is a compact operator from $C(D) \to C(D)$.

Still for the case $D = [a, b]$, another way to show that $k(x, y)$ satisfies (A_1) and (A_2) is to rewrite k in the form

$$k(x, y) = \sum_{i=0}^{n} h_i(x, y) l_i(x, y) \tag{2.8.9}$$

for some $n > 0$, with each $l_i(x, y)$ continuous for $a \le x, y \le b$ and each $h_i(x, y)$ satisfying (A_1)–(A_2). It is left to the reader to show that in this case, k also satisfies (A_1)–(A_2). The utility of this approach is that it is sometimes difficult to show directly that k satisfies (A_1)–(A_2), whereas showing (2.8.9) with h_i, l_i satisfying the specified conditions may be easier.

Example 2.8.2 Let $[a, b] = [0, \pi]$ and $k(x, y) = \log|\cos x - \cos y|$. Rewrite the kernel function as

$$k(x, y) = \underbrace{|x - y|^{-1/2}}_{h(x,y)} \underbrace{|x - y|^{1/2} \log|\cos x - \cos y|}_{l(x,y)}. \tag{2.8.10}$$

Easily, l is continuous. From the discussion following (2.8.8), h satisfies (A_1)–(A_2). Thus k is the kernel of a compact integral operator on $C[0, \pi]$ to $C[0, \pi]$. □

2.8.2 Properties of compact operators

Another way of obtaining compact operators is to look at limits of simpler "finite-dimensional operators" in $\mathcal{L}(V, W)$, the Banach space of bounded linear operators from V to W. This gives another perspective on compact operators, one that leads to improved intuition by emphasizing their close relationship to operators on finite dimensional spaces.

Definition 2.8.3 *Let V and W be linear spaces. The linear operator $K : V \to W$ is of* finite rank *if $\mathcal{R}(K)$, the range of K, is finite dimensional.*

Proposition 2.8.4 *Let V and W be normed spaces, and let $K : V \to W$ be a bounded finite rank operator. Then K is a compact operator.*

Proof. The range $\mathcal{R}(K)$ is a finite-dimensional normed space, and therefore it is complete. Consider the set

$$S = \{Kv \mid \|v\|_V \leq 1\}.$$

It is bounded, each of its elements being bounded by $\|K\|$. Notice that $S \subset \mathcal{R}(K)$. Then S has a compact closure, since all bounded closed sets in a finite dimensional space are compact. This shows K is compact. □

Example 2.8.5 Let $V = W = C[a, b]$ with the norm $\|\cdot\|_\infty$. Consider the kernel function

$$k(x, y) = \sum_{i=1}^{n} \beta_i(x)\, \gamma_i(y) \tag{2.8.11}$$

with each β_i continuous on $[a, b]$ and each γ_i absolutely integrable on $[a, b]$. Then the associated integral operator K is a bounded, finite rank operator on $C[a, b]$ to $C[a, b]$:

$$Kv(x) = \sum_{i=1}^{n} \beta_i(x) \int_a^b \gamma_i(y)\, v(y)\, dy, \quad v \in C[a, b]. \tag{2.8.12}$$

Indeed, we have

$$\|K\| \leq \sum_{i=1}^{n} \|\beta_i\|_\infty \int_a^b |\gamma_i(y)|\, dy.$$

From (2.8.12), $Kv \in C[a, b]$ and $\mathcal{R}(K) \subset \text{span}\{\beta_1, \ldots, \beta_n\}$, a finite dimensional space. □

Kernel functions of the form (2.8.11) are called *degenerate*. Below we see that the associated integral equation $(\lambda I - K)v = f$, $\lambda \neq 0$, is essentially a finite dimensional equation.

Proposition 2.8.6 *Let $K \in \mathcal{L}(U, V)$ and $L \in \mathcal{L}(V, W)$ with at least one of them being compact. Then LK is a compact operator from U to W.*

The proof is left as Exercise 2.8.1 for the reader.

The following result gives the framework for using finite rank operators to obtain similar, but more general compact operators.

Proposition 2.8.7 *Let V and W be normed spaces, with W complete. Assume $\{K_n\} \subset \mathcal{L}(V, W)$ is a sequence of compact operators such that $K_n \to K$ in $\mathcal{L}(V, W)$. Then K is compact.*

This is a standard result found in most books on functional analysis; e.g. see [58, p. 174] or [71, p. 486].

For almost all function spaces V that occur in applied mathematics, the compact operators can be characterized as being the limit of a sequence

of bounded finite-rank operators. This gives a further justification for the presentation of Proposition 2.8.7.

Example 2.8.8 Let D be a closed and bounded set in $\mathbb{R}^d$. For example, D could be a region with nonempty interior, a piecewise smooth surface, or a piecewise smooth curve. Let $k(\boldsymbol{x}, \boldsymbol{y})$ be a continuous function of $\boldsymbol{x}, \boldsymbol{y} \in D$. Suppose we can define a sequence of continuous degenerate kernel functions $k_n(\boldsymbol{x}, \boldsymbol{y})$ for which

$$\max_{\boldsymbol{x} \in D} \int_D |k(\boldsymbol{x}, \boldsymbol{y}) - k_n(\boldsymbol{x}, \boldsymbol{y})| \, dy \to 0 \quad \text{as } n \to \infty. \tag{2.8.13}$$

Then for the associated integral operators, it easily follows that $K_n \to K$; and by Proposition 2.8.7, K is compact. The condition (2.8.13) is true for general continuous functions $k(x, y)$, and we leave to the exercises the proof for various choices of D. Of course, we already knew that K was compact in this case, from the discussion following (2.8.8). But the present approach shows the close relationship of compact operators and finite dimensional operators. □

Example 2.8.9 Let $V = W = C[a, b]$ with the norm $\|\cdot\|_\infty$. Consider the kernel function

$$k(x, y) = \frac{1}{|x - y|^\gamma} \tag{2.8.14}$$

for some $0 < \gamma < 1$. Define a sequence of continuous kernel functions to approximate it:

$$k_n(x, y) = \begin{cases} \dfrac{1}{|x - y|^\gamma}, & |x - y| \geq \dfrac{1}{n}, \\ n^\gamma, & |x - y| \leq \dfrac{1}{n}. \end{cases} \tag{2.8.15}$$

This merely limits the height of the graph of $k_n(x, y)$ to that of $k(x, y)$ when $|x - y| = 1/n$. Easily, $k_n(x, y)$ is a continuous function for $a \leq x, y \leq b$, and thus the associated integral operator K_n is compact on $C[a, b]$. For the associated integral operators,

$$\|K - K_n\| = \frac{2\gamma}{1 - \gamma} \cdot \frac{1}{n^{1-\gamma}}$$

which converges to zero as $n \to \infty$. By Proposition 2.8.7, K is a compact operator on $C[a, b]$. □

2.8.3 Integral operators on $L^2(a, b)$

Let $V = W = L^2(a, b)$, and let K be the integral operator associated with a kernel function $k(x, y)$. We first show that under suitable assumptions on

k, the operator K maps $L^2(a,b)$ to $L^2(a,b)$ and is bounded. Assume

$$M \equiv \left[\int_a^b \int_a^b |k(x,y)|^2 dy\, dx\right]^{1/2} < \infty. \tag{2.8.16}$$

For $v \in L^2(a,b)$, we use the Cauchy-Schwarz inequality to obtain

$$\begin{aligned}\|Kv\|_2^2 &= \int_a^b \left|\int_a^b k(x,y)v(y)dy\right|^2 dx \\ &\le \int_a^b \left[\int_a^b |K(x,y)|^2 dy\right]\left[\int_a^b |v(y)|^2 dy\right] dx \\ &= M^2\|v\|_2^2.\end{aligned}$$

Thus, $Kv \in L^2(a,b)$ and

$$\|K\| \le M. \tag{2.8.17}$$

This bound is comparable to the use of the Frobenius matrix norm to bound the operator norm of a matrix A: $\mathbb{R}^n \to \mathbb{R}^m$, when the vector norm $\|\cdot\|_2$ is being used. Recall that the Frobenius norm of a matrix A is given by

$$\|A\|_F = \Big(\sum_{i,j} |A_{i,j}|^2\Big)^{1/2}.$$

Kernel functions k for which $M < \infty$ are called *Hilbert-Schmidt kernel functions*, and the quantity M in (2.8.16) is called the *Hilbert-Schmidt norm* of K.

For integral operators K with a degenerate kernel function as in (2.8.11), the operator K is bounded if all β_i, $\gamma_i \in L^2(a,b)$. This is a straightforward result which we leave as a problem for the reader. From Proposition 2.8.4, the integral operator is then compact.

To examine the compactness of K for more general kernel functions, we assume there is a sequence of kernel functions $k_n(x,y)$ for which (i) $K_n : L^2(a,b) \to L^2(a,b)$ is compact, and (ii)

$$M_n \equiv \left[\int_a^b \int_a^b |k(x,y) - k_n(x,y)|^2 dy\, dx\right]^{1/2} \to 0 \quad \text{as } n \to \infty. \tag{2.8.18}$$

For example, if K is continuous, then this follows from (2.8.13). The operator $K - K_n$ is an integral operator, and we apply (2.8.16)–(2.8.17) to it to obtain

$$\|K - K_n\| \le M_n \to 0 \quad \text{as } n \to \infty.$$

From Proposition 2.8.7, this shows K is compact. For any Hilbert-Schmidt kernel function, (2.8.18) can be shown to hold for a suitable choice of degenerate kernel functions k_n.

We leave it for the reader to show that $\log|x-y|$ and $|x-y|^{-\gamma}$, $\gamma < \frac{1}{2}$, are Hilbert-Schmidt kernel functions (Exercise 2.8.4). For $\frac{1}{2} \leq \gamma < 1$, the kernel function $|x-y|^{-\gamma}$ still defines a compact integral operator K on $L^2(a,b)$, but the above theory for Hilbert-Schmidt kernel functions does not apply. For a proof of the compactness of K in this case, see Mikhlin [172, p. 160].

2.8.4 The Fredholm alternative theorem

Integral equations were studied in the 19th century as one means of investigating boundary value problems for the Laplace equation, for example,

$$\begin{aligned} \Delta u(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \Omega, \\ u(\boldsymbol{x}) &= f(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega, \end{aligned} \tag{2.8.19}$$

and other elliptic partial differential equations. In the early 1900's, Ivar Fredholm found necessary and sufficient conditions for the solvability of a large class of Fredholm integral equations of the second kind. With these results, he then was able to give much more general existence theorems for the solution of boundary value problems such as (2.8.19). In this subsection, we state and prove the most important results of Fredholm; and in the following subsection, we give additional results without proof.

The theory of integral equations has been due to many people, with David Hilbert being among the most important popularizers of the area. The subject of integral equations continues as an important area of study in applied mathematics; and for an introduction that includes a review of much recent literature, see Kress [149]. For an interesting historical account of the development of functional analysis as it was affected by the development of the theory of integral equations, see Bernkopf [36]. From hereon, to simplify notation, for a scalar λ and an operator $K : V \to V$, we use $\lambda - K$ for the operator $\lambda I - K$, where $I : V \to V$ is the identity operator.

Theorem 2.8.10 (Fredholm alternative) *Let V be a Banach space, and let $K : V \to V$ be compact. Then the equation $(\lambda - K)u = f$, $\lambda \neq 0$, has a unique solution $u \in V$ for any $f \in V$ if and only if the homogeneous equation $(\lambda - K)v = 0$ has only the trivial solution $v = 0$. In such a case, the operator $\lambda - K : V \overset{1-1}{\underset{onto}{\to}} V$ has a bounded inverse $(\lambda - K)^{-1}$.*

Proof. The theorem is true for any compact operator K, but here we give a proof only for those compact operators which are the limit of a sequence of bounded finite-rank operators. For a more general proof, see Kress [149, Chap. 3] or Conway [58, p. 217]. We remark that the theorem is a generalization of the following standard result for finite dimensional vector spaces V. For A a matrix of order n, with $V = \mathbb{R}^n$ or $\mathbb{C}^n$ (with A having real entries for the former case), the linear system $Au = w$ has a

unique solution $u \in V$ for any $w \in V$ if and only if the homogeneous linear system $Az = 0$ has only the zero solution $z = 0$.
(a) We begin with the case where K is finite-rank and bounded. Let $\{\varphi_1, \ldots, \varphi_n\}$ be a basis for $\mathcal{R}(K)$, the range of K. Rewrite the equation $(\lambda - K)u = f$ as

$$u = \frac{1}{\lambda}(f + Ku). \tag{2.8.20}$$

If this equation has a unique solution $u \in V$, then

$$u = \frac{1}{\lambda}(f + c_1\varphi_1 + \cdots + c_n\varphi_n) \tag{2.8.21}$$

for some uniquely determined set of constants $c_1, \ldots, c_n$.

By substituting (2.8.21) into the equation $(\lambda - K)\,u = f$, we have

$$\lambda\left(\frac{1}{\lambda}f + \frac{1}{\lambda}\sum_{i=1}^{n} c_i\varphi_i\right) - \frac{1}{\lambda}Kf - \frac{1}{\lambda}\sum_{j=1}^{n} c_jK\varphi_j = f.$$

Multiply by λ, and then simplify to obtain

$$\lambda\sum_{i=1}^{n} c_i\varphi_i - \sum_{j=1}^{n} c_jK\varphi_j = Kf. \tag{2.8.22}$$

Using the basis $\{\varphi_i\}$ for $\mathcal{R}(K)$, write

$$Kf = \sum_{i=1}^{n} \gamma_i\varphi_i$$

and

$$K\varphi_j = \sum_{i=1}^{n} a_{ij}\varphi_i, \quad 1 \le j \le n.$$

The coefficients $\{\gamma_i\}$ and $\{a_{ij}\}$ are uniquely determined. Substitute these expressions into (2.8.22) and rearrange the terms,

$$\sum_{i=1}^{n}\left(\lambda c_i - \sum_{j=1}^{n} a_{ij}c_j\right)\varphi_i = \sum_{i=1}^{n} \gamma_i\varphi_i.$$

By the independence of the basis elements φ_i, we obtain the linear system

$$\lambda c_i - \sum_{j=1}^{n} a_{ij}c_j = \gamma_i, \quad 1 \le i \le n. \tag{2.8.23}$$

Claim: This linear system and the equation $(\lambda - K)u = f$ are completely equivalent in their solvability, with (2.8.21) furnishing a one-to-one correspondence between the solutions of the two of them.

We have shown above that if u is a solution of $(\lambda - K)u = f$, then $(c_1, \ldots, c_n)^T$ is a solution of (2.8.23). In addition, suppose u_1 and u_2 are distinct solutions of $(\lambda - K)u = f$. Then

$$Ku_1 = \lambda u_1 - f \quad \text{and} \quad Ku_2 = \lambda u_2 - f, \quad \lambda \neq 0,$$

are also distinct vectors in $\mathcal{R}(K)$, and thus the associated vectors of coordinates $(c_1^{(1)}, \ldots, c_n^{(1)})^T$ and $(c_1^{(2)}, \ldots, c_n^{(2)})^T$,

$$K\varphi_i = \sum_{k=1}^{n} c_k^{(i)} \varphi_k, \quad i = 1, 2$$

must also be distinct.

For the converse statement, suppose $(c_1, \ldots, c_n)^T$ is a solution of (2.8.23). Define a vector $u \in V$ by using (2.8.21), and then check whether this u satisfies the integral equation:

$$\begin{aligned}
(\lambda - K)u &= \lambda \left(\frac{1}{\lambda} f + \frac{1}{\lambda} \sum_{i=1}^{n} c_i \varphi_i \right) - \frac{1}{\lambda} Kf - \frac{1}{\lambda} \sum_{j=1}^{n} c_j K\varphi_j \\
&= f + \frac{1}{\lambda} \left(\lambda \sum_{i=1}^{n} c_i \varphi_i - Kf - \sum_{j=1}^{n} c_j K\varphi_j \right) \\
&= f + \frac{1}{\lambda} \left(\sum_{i=1}^{n} \lambda c_i \varphi_i - \sum_{i=1}^{n} \gamma_i \varphi_i - \sum_{j=1}^{n} c_j \sum_{i=1}^{n} a_{ij} \varphi_i \right) \\
&= f + \frac{1}{\lambda} \sum_{i=1}^{n} \underbrace{\left(\lambda c_i - \gamma_i - \sum_{j=1}^{n} a_{ij} c_j \right)}_{=0,\ i=1,\ldots,n} \varphi_i \\
&= f.
\end{aligned}$$

Also, distinct coordinate vectors $(c_1, \ldots, c_n)^T$ lead to distinct solutions u given by (2.8.21), because of the linear independence of the basis vectors $\{\varphi_1, \ldots, \varphi_n\}$. This completes the proof of the claim given above.

Now consider the Fredholm alternative theorem for $(\lambda - K)u = f$ with this finite rank operator K. Suppose

$$\lambda - K : V \underset{onto}{\overset{1-1}{\to}} V.$$

Then trivially, the null space $\mathcal{N}(\lambda - K) = \{0\}$. For the converse, assume $(\lambda - K)\, v = 0$ has only the solution $v = 0$. Note that we want to show that $(\lambda - K)\, u = f$ has a unique solution for every $f \in V$.

Consider the associated linear system (2.8.23). It can be shown to have a unique solution for any right hand side $(\gamma_1, \ldots, \gamma_n)^T$ by showing that

the homogeneous linear system has only the zero solution. The latter is done by means of the equivalence of the homogeneous linear system to the homogeneous equation $(\lambda - K)v = 0$, which implies $v = 0$. But since (2.8.23) has a unique solution, so must $(\lambda - K)u = f$, and it is given by (2.8.21).

We must also show that $(\lambda - K)^{-1}$ is bounded. This can be done directly by a further examination of the consequences of K being a bounded and finite rank operator; but it is simpler to just cite the open mapping theorem (see Theorem 2.4.3).

(b) Assume now that $\|K - K_n\| \to 0$, with K_n finite rank and bounded. Rewrite $(\lambda - K)\, u = f$ as

$$[\lambda - (K - K_n)]\, u = f + K_n u, \quad n \geq 1. \tag{2.8.24}$$

Pick an index $m > 0$ for which

$$\|K - K_m\| < |\lambda| \tag{2.8.25}$$

and fix it. By the geometric series theorem (Theorem 2.3.1),

$$Q_m \equiv [\lambda - (K - K_m)]^{-1}$$

exists and is bounded, with

$$\|Q_m\| \leq \frac{1}{|\lambda| - \|K - K_m\|}.$$

The equation (2.8.24) with $n = m$ can now be written in the equivalent form

$$u - Q_m K_m u = Q_m f. \tag{2.8.26}$$

The operator $Q_m K_m$ is bounded and finite rank. The boundedness follows from that of Q_m and K_m. To show it is finite rank, let $\mathcal{R}(K_m) = \text{span}\{\varphi_1, \ldots, u_m\}$. Then

$$\mathcal{R}(Q_m K_m) = \text{span}\{Q_m \varphi_1, \ldots, Q_m u_m\}$$

is a finite-dimensional space.

The equation (2.8.26) is one to which we can apply part (a) of this proof. Assume $(\lambda - K)v = 0$ implies $v = 0$. By the above equivalence, this yields

$$(I - Q_m K_m)\, v = 0 \quad \Longrightarrow \quad v = 0.$$

But from part (a), this says $(I - Q_m K_m)\, u = w$ has a unique solution u for every $w \in V$, and in particular, for $w = Q_m f$ as in (2.8.26). By the equivalence of (2.8.26) and $(\lambda - K)\, u = f$, we have that the latter is uniquely solvable for every $f \in V$. The boundedness of $(\lambda - K)^{-1}$ follows

from part (a) and the boundedness of Q_m. Alternatively, the open mapping theorem can be cited, as earlier in part (a). □

For many practical problems in which K is not compact, it is important to note what makes this proof work. It is *not* necessary to have a sequence of bounded and finite rank operators $\{K_n\}$ for which $\|K - K_n\| \to 0$. Rather, it is necessary to satisfy the inequality (2.8.25) for one finite rank operator K_m; and in applying the proof to other operators K, it is necessary only that K_m be compact. In such a case, the proof following (2.8.25) remains valid, and the Fredholm Alternative still applies to such an equation

$$(\lambda - K)\, u = f.$$

2.8.5 Additional results on Fredholm integral equations

In this subsection, we give additional results on the solvability of compact equations of the second kind, $(\lambda - K)\, u = f$, with $\lambda \neq 0$. No proofs are given for these results, and the reader is referred to a standard text on integral equations; e.g. see Kress [149] or Mikhlin [171].

Definition 2.8.11 *Let $K : V \to V$. If there is a scalar λ and an associated vector $u \neq 0$ for which $Ku = \lambda u$, then λ is called an* eigenvalue *and u an associated* eigenvector *of the operator K.*

When dealing with compact operators K, we generally are interested in only the nonzero eigenvalues of K.

In the following, recall that $\mathcal{N}(A)$ denotes the null space of A.

Theorem 2.8.12 *Let $K : V \to V$ be compact, and let V be a Banach space. Then:*

(1) *The eigenvalues of K form a discrete set in the complex plane $\mathbb{C}$, with 0 as the only possible limit point.*

(2) *For each nonzero eigenvalue λ of K, there are only a finite number of linearly independent eigenvectors.*

(3) *Each nonzero eigenvalue λ of K has finite index $\nu(\lambda) \geq 1$. This means*

$$\begin{aligned}\mathcal{N}(\lambda - K) &\underset{\neq}{\subset} \mathcal{N}((\lambda - K)^2) \underset{\neq}{\subset} \cdots \\ &\underset{\neq}{\subset} \mathcal{N}((\lambda - K)^{\nu(\lambda)}) = \mathcal{N}((\lambda - K)^{\nu(\lambda)+1}).\end{aligned} \tag{2.8.27}$$

In addition, $\mathcal{N}((\lambda - K)^{\nu(\lambda)})$ is finite dimensional. The elements of the subspace $\mathcal{N}((\lambda - K)^{\nu(\lambda)}) \backslash \mathcal{N}(\lambda - K)$ are called generalized eigenvectors of K.

(4) *For any* $\lambda \neq 0$, $\mathcal{R}(\lambda - K)$ *is closed in* V.

(5) *For each nonzero eigenvalue* λ *of* K,

$$V = \mathcal{N}((\lambda - K)^{\nu(\lambda)}) \oplus \mathcal{R}((\lambda - K)^{\nu(\lambda)}) \tag{2.8.28}$$

is a decomposition of V *into invariant subspaces. This implies that every* $u \in V$ *can be written as* $u = u_1 + u_2$ *with unique choices*

$$u_1 \in \mathcal{N}((\lambda - K)^{\nu(\lambda)}) \quad \text{and} \quad u_2 \in \mathcal{R}((\lambda - K)^{\nu(\lambda)}).$$

Being invariant means that

$$K : \mathcal{N}((\lambda - K)^{\nu(\lambda)}) \to \mathcal{N}((\lambda - K)^{\nu(\lambda)}),$$
$$K : \mathcal{R}((\lambda - K)^{\nu(\lambda)}) \to \mathcal{R}((\lambda - K)^{\nu(\lambda)}).$$

(6) *The Fredholm alternative theorem and the above results* (1)–(5) *remain true if* K^m *is compact for some integer* $m > 1$.

For results on the speed with which the eigenvalues $\{\lambda_n\}$ of compact integral operators K converge to zero, see Hille and Tamarkin [125] and Fenyö and Stolle [80, Section 8.9]. Generally, as the differentiability of the kernel function $k(x, y)$ increases, the speed of convergence to zero of the eigenvalues also increases.

For the following results, recall from Section 2.6 the concept of an adjoint operator.

Lemma 2.8.13 *Let* V *be a Hilbert space with scalars the complex numbers* $\mathbb{C}$*, and let* $K : V \to V$ *be a compact operator. Then* $K^* : V \to V$ *is also a compact operator.*

This implies that the operator K^* also shares the properties stated above for the compact operator K. There is, however, a closer relationship between the operators K and K^*, which is given in the following theorem.

Theorem 2.8.14 *Let* V *be a Hilbert space with scalars the complex numbers* $\mathbb{C}$*, let* $K : V \to V$ *be a compact operator, and let* λ *be a nonzero eigenvalue of* K*. Then:*

(1) $\bar{\lambda}$ *is an eigenvalue of the adjoint operator* K^**. In addition,* $\mathcal{N}(\lambda - K)$ *and* $\mathcal{N}(\bar{\lambda} - K^*)$ *have the same dimension.*

(2) *The equation* $(\lambda - K)u = f$ *is solvable if and only if*

$$(f, v) = 0 \quad \forall\, v \in \mathcal{N}(\bar{\lambda} - K^*). \tag{2.8.29}$$

An equivalent way of writing this is

$$\mathcal{R}(\lambda - K) = \mathcal{N}(\bar{\lambda} - K^*)^{\perp},$$

the subspace orthogonal to $\mathcal{N}(\bar{\lambda} - K^*)$. *With this, we can write the decomposition*

$$V = \mathcal{N}(\bar{\lambda} - K^*) \oplus \mathcal{R}(\lambda - K). \tag{2.8.30}$$

Theorem 2.8.15 *Let* V *be a Hilbert space with scalars the complex numbers* $\mathbb{C}$, *and let* $K : V \to V$ *be a self-adjoint compact operator. Then all eigenvalues of* K *are real and of index* $\nu(\lambda) = 1$. *In addition, the corresponding eigenvectors can be chosen to form an orthonormal set. Order the nonzero eigenvalues as follows:*

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \geq \cdots > 0 \tag{2.8.31}$$

with each eigenvalue repeated according to its multiplicity (i.e. the dimension of $\mathcal{N}(\lambda - K)$*). Then we write*

$$Ku_i = \lambda_i u_i, \quad i \geq 1 \tag{2.8.32}$$

with

$$(u_i, u_j) = \delta_{ij}.$$

Also, the eigenvectors $\{u_i\}$ *form an orthonormal basis for* $\overline{\mathcal{R}(\lambda - K)}$.

Much of the theory of self-adjoint boundary value problems for ordinary and partial differential equations is based on Theorems 2.8.14 and 2.8.15. Moreover, the completeness in $L^2(D)$ of many families of functions is proven by showing they are the eigenfunctions to a self-adjoint differential equation or integral equation problem.

Example 2.8.16 Let $D = \{\boldsymbol{x} \in \mathbb{R}^3 \mid \|\boldsymbol{x}\| = 1\}$, the unit sphere in $\mathbb{R}^3$, and let $V = L^2(D)$. Here $\|\boldsymbol{x}\|$ denotes the Euclidean length of $\boldsymbol{x}$. Define

$$Kv(\boldsymbol{x}) = \int_D \frac{v(\boldsymbol{y})}{\|\boldsymbol{x} - \boldsymbol{y}\|} dS_{\boldsymbol{y}}, \quad \boldsymbol{x} \in D. \tag{2.8.33}$$

This is a compact operator, a proof of which is given in Mikhlin [172, p. 160]. The eigenfunctions of K are called *spherical harmonics*, a much-studied set of functions; e.g. see [86], [161]. For each integer $k \geq 0$, there are $2k + 1$ independent spherical harmonics of *degree* k; and for each such spherical harmonic φ_k, we have

$$K\varphi_k = \frac{4\pi}{2k+1}\varphi_k, \quad k = 0, 1, \ldots. \tag{2.8.34}$$

Letting $\mu_k = 4\pi/(2k+1)$, we have $\mathcal{N}(\mu_k - K)$ has dimension $2k+1$, $k \geq 0$. It is well-known that the set of all spherical harmonics forms a basis for $L^2(D)$, in agreement with Theorem 2.8.15. □

Exercise 2.8.1 Prove Proposition 2.8.6.

Exercise 2.8.2 Suppose k is a degenerate kernel function given by (2.8.11) with all $\beta_i, \gamma_i \in L^2(a,b)$. Show that the integral operator K, defined by

$$Kv(x) = \int_a^b k(x,y)\, v(y)\, dy$$

is bounded from $L^2(a,b)$ to $L^2(a,b)$.

Exercise 2.8.3 Consider the integral operator (2.8.2). Assume the kernel function k has the form (2.8.9) with each $l_i(x,y)$ continuous for $a \le x, y \le b$ and each $h_i(x,y)$ satisfying (A_1)–(A_2). Prove that k also satisfies (A_1)–(A_2).

Exercise 2.8.4 Show that $\log|x-y|$ and $|x-y|^{-\gamma}$, $\gamma < \frac{1}{2}$, are Hilbert-Schmidt kernel functions.

Exercise 2.8.5 Consider the integral equation

$$\lambda f(x) - \int_0^1 e^{x-y} f(y)\, dy = g(x), \quad 0 \le x \le 1$$

with $g \in C[0,1]$. Denote the integral operator in the equation by K. Consider K as a mapping on $C[0,1]$ into itself, and use the uniform norm $\|\cdot\|_\infty$. Find a bound for the condition number

$$\operatorname{cond}(\lambda - K) \equiv \|\lambda - K\|\, \|(\lambda - K)^{-1}\|$$

within the framework of the space $C[0,1]$. Do this for all values of λ for which $(\lambda - K)^{-1}$ exists as a bounded operator on $C[0,1]$. Comment on how the condition number varies with λ.

Exercise 2.8.6 Similar to Example 2.3.6 of Section 2.3, use the approximation

$$e^{xy} \approx 1 + xy$$

to examine the solvability of the integral equation

$$\lambda u(x) - \int_0^1 e^{xy} u(y)\, dy = f(x), \quad 0 \le x \le 1.$$

To solve the integral equation associated with the kernel $1 + xy$, use the method developed in the proof of Theorem 2.8.10.

Exercise 2.8.7 For any $f \in C[0,1]$, define

$$\mathcal{A}f(x) = \begin{cases} \displaystyle\int_0^x \frac{f(y)}{\sqrt{x^2 - y^2}} dy, & 0 < x \le 1, \\ \displaystyle\frac{\pi}{2} f(0), & x = 0. \end{cases}$$

This is called an *Abel integral operator*. Show that $f(x) = x^\alpha$ is an eigenfunction of $\mathcal{A}$ for every $\alpha \ge 0$. What is the corresponding eigenvalue? Can $\mathcal{A}$ be a compact operator?

2.9 The resolvent operator

Let V be a complex Banach space, e.g. $V = C(D)$ the set of continuous complex-valued functions on a closed set D with the uniform norm $\|\cdot\|_\infty$; and let $L : V \to V$ be a bounded linear operator. From the geometric series theorem, Theorem 2.3.1, we know that if $|\lambda| > \|L\|$, then $(\lambda - L)^{-1}$ exists as a bounded linear operator from V to V. It is useful to consider the set of all complex numbers λ for which such an inverse operator $(\lambda - L)^{-1}$ exists on V to V.

Definition 2.9.1 (a) *Let V be a complex Banach space, and let $L : V \to V$ be a bounded linear operator. We say $\lambda \in \mathbb{C}$ belongs to the* resolvent set *of L if $(\lambda - L)^{-1}$ exists as a bounded linear operator from V to V. The resolvent set of L is denoted by $\rho(L)$. The operator $(\lambda - L)^{-1}$ is called the* resolvent operator.
(b) *The set $\sigma(L) = \mathbb{C} \backslash \rho(L)$ is called the* spectrum *of L.*

From the remarks preceding the definition,

$$\{\lambda \in \mathbb{C} \mid |\lambda| > \|L\|\} \subset \rho(L).$$

In addition, we have the following.

Lemma 2.9.2 *The set $\rho(L)$ is open in $\mathbb{C}$; and consequently, $\sigma(L)$ is a closed set.*

Proof. Let $\lambda_0 \in \rho(L)$. We use the perturbation result in Theorem 2.3.5 to show that all points λ in a sufficiently small neighborhood of λ_0 are also in $\rho(L)$; this is sufficient for showing $\rho(L)$ is open. Since $(\lambda_0 - L)^{-1}$ is a bounded linear operator on V to V, consider all $\lambda \in \mathbb{C}$ for which

$$|\lambda - \lambda_0| < \frac{1}{\|(\lambda_0 - L)^{-1}\|}. \tag{2.9.1}$$

Using Theorem 2.3.5, we have that $(\lambda - L)^{-1}$ exists as a bounded operator from V to V, and moreover,

$$\left\|(\lambda - L)^{-1} - (\lambda_0 - L)^{-1}\right\| \leq \frac{|\lambda - \lambda_0| \left\|(\lambda_0 - L)^{-1}\right\|^2}{1 - |\lambda - \lambda_0| \left\|(\lambda_0 - L)^{-1}\right\|}. \tag{2.9.2}$$

This shows

$$\{\lambda \in \mathbb{C} \mid |\lambda - \lambda_0| < \varepsilon\} \subset \rho(L)$$

provided ε is chosen sufficiently small. Hence, $\rho(L)$ is an open set.

The inequality (2.9.2) shows that $R(\lambda) \equiv (\lambda - L)^{-1}$ is a continuous function of λ from $\mathbb{C}$ to $\mathcal{L}(V)$. □

A complex number λ can belong to $\sigma(L)$ for several different reasons. Following is a standard classification scheme.

1. *Point spectrum.* $\lambda \in \sigma_P(L)$ means that λ is an eigenvalue of L. Thus there is a nonzero eigenvector $u \in V$ for which $Lu = \lambda u$. Such cases were explored in Section 2.8 with L a compact operator. In this case, the nonzero portion of $\sigma(L)$ consists entirely of eigenvalues, and moreover, 0 is the only possible point in $\mathbb{C}$ to which sequences of eigenvalues can converge.

2. *Continuous spectrum.* $\lambda \in \sigma_C(L)$ means that $(\lambda - L)$ is one-to-one, $\mathcal{R}(\lambda - L) \neq V$, and $\overline{\mathcal{R}(\lambda - L)} = V$. Note that if $\lambda \neq 0$, then L cannot be compact. (Why?) This type of situation, $\lambda \in \sigma_C(L)$, occurs in solving equations $(\lambda - L)u = f$ that are ill-posed. In the case $\lambda = 0$, such equations can often be written as an integral equation of the first kind
$$\int_a^b \ell(x, y)u(y)\, dy = f(x), \quad a \leq x \leq b,$$
with $\ell(x, y)$ continuous and smooth.

3. *Residual spectrum.* $\lambda \in \sigma_R(L)$ means $\lambda \in \sigma(L)$ and that it is in neither the point spectrum nor continuous spectrum. This case can be further subdivided, into cases with $\mathcal{R}(\lambda - L)$ closed and not closed. The latter case consists of ill-posed problems, much as with the case of continuous spectrum. For the former case, the equation $(\lambda - L)u = f$ is usually a well-posed problem; but some change in it is often needed when developing practical methods of solution.

If L is a compact operator on V to V, and if V is infinite dimensional, then it can be shown that $0 \in \sigma(L)$. In addition in this case, if 0 is not an eigenvalue of L, then L^{-1} can be shown to be unbounded on $\mathcal{R}(L)$. Equations $Lu = f$ with L compact make up a significant proportion of ill-posed problems.

2.9.1 $R(\lambda)$ as a holomorphic function

Let $\lambda_0 \in \rho(L)$. Returning to the proof of Lemma 2.9.2, we can write $R(\lambda) \equiv (\lambda - L)^{-1}$ as
$$R(\lambda) = \sum_{k=0}^{\infty} (-1)^k (\lambda - \lambda_0)^k R(\lambda_0)^{k+1} \tag{2.9.3}$$
for all λ satisfying (2.9.1). Thus we have a power series expansion of $R(\lambda)$ about the point λ_0. This can be used to introduce the notion that R is an analytic (or holomorphic) function from $\rho(L) \subset \mathbb{C}$ to the vector space $\mathcal{L}(V)$. Many of the definitions, ideas, and results of *complex analysis* can be extended to analytic operator-valued functions. See [71, p. 566] for an introduction to these ideas.

In particular, we can introduce line integrals. We are especially interested in line integrals of the form

$$g_\Gamma(L) = \frac{1}{2\pi i}\int_\Gamma (\mu - L)^{-1} g(\mu)\, d\mu. \tag{2.9.4}$$

Note that whereas $g : \rho(L) \to \mathbb{C}$, the quantity $g_\Gamma(L) \in \mathcal{L}(V)$. In this integral, Γ is a piecewise smooth curve of finite length in $\rho(L)$; and Γ can consist of several finite disjoint curves. In complex analysis, such integrals occur in connection with studying *Cauchy's theorem*.

Let $\mathcal{F}(L)$ denote the set of all functions g which are analytic on some open set U containing $\sigma(L)$, with the set U dependent on the function g (U need not be connected). For functions in $\mathcal{F}(L)$, a number of important results can be shown for the operators $g(L)$ of (2.9.4) with $g \in \mathcal{F}(L)$. For a proof of the following, see [71, p. 568].

Theorem 2.9.3 *Let $f, g \in \mathcal{F}(L)$, and let $f_\Gamma(L)$, $g_\Gamma(L)$ be defined using* (2.9.4)*, assuming Γ is located within the domain of analyticity of both f and g. Then*
(a) *$f \cdot g \in \mathcal{F}(L)$, and $f_\Gamma(L) \cdot g_\Gamma(L) = (f \cdot g)_\Gamma(L)$;*
(b) *if f has a power series expansion*

$$f(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n$$

that is valid in some open disk about $\sigma(L)$, then

$$f_\Gamma(L) = \sum_{n=0}^{\infty} a_n L^n.$$

In numerical analysis, such integrals (2.9.4) become a means for studying the convergence of algorithms for approximating the eigenvalues of L.

Theorem 2.9.4 *Let L be a compact operator from V to V, and let λ_0 be a nonzero eigenvalue of L. Introduce*

$$E(\lambda_0, L) = \frac{1}{2\pi i}\int_{|\lambda - \lambda_0| = \varepsilon} (\lambda - L)^{-1}\, d\lambda \tag{2.9.5}$$

with ε less than the distance from λ_0 to the remaining portion of $\sigma(L)$. Then:
(a) *$E(\lambda_0, L)$ is a projection operator on V to V.*
(b) *$E(\lambda_0, L)V$ is the set of all ordinary and generalized eigenvectors associated with λ_0, i.e.*

$$E(\lambda_0, L)V = \mathcal{N}((\lambda - K)^{\nu(\lambda_0)})$$

with the latter taken from (2.8.27) *and $\nu(\lambda_0)$ the index of λ_0.*

For a proof of these results, see Dunford and Schwartz [71, pp. 566–580].

When L is approximated by a sequence of operators $\{L_n\}$, we can examine the convergence of the eigenspaces of L_n to those of L by means of tools fashioned from (2.9.5). Examples of such analyses can be found in [11], [13], and Chatelin [48].

Exercise 2.9.1 Let $\lambda \in \rho(L)$. Define $d(\lambda)$ to be the distance from λ to $\sigma(L)$,

$$d(\lambda) = \min_{\kappa \in \sigma(L)} |\lambda - \kappa| .$$

Show that

$$\left\|(\lambda - L)^{-1}\right\| \geq \frac{1}{d(\lambda)}.$$

This implies $\left\|(\lambda - L)^{-1}\right\| \to \infty$ as $\lambda \to \sigma(L)$.

Exercise 2.9.2 Let $V = C[0,1]$, and let L be the Volterra integral operator

$$Lv(x) = \int_0^x k(x,y)\, v(y)\, dy, \quad 0 \leq x \leq 1,\ v \in C[0,1]$$

with $k(x,y)$ continuous for $0 \leq y \leq x \leq 1$. What is $\sigma(L)$?

Exercise 2.9.3 Derive the formula (2.9.3).

Exercise 2.9.4 Let $F \subset \rho(L)$ be closed and bounded. Show $(\lambda - L)^{-1}$ is a continuous function of $\lambda \in F$, with

$$\max_{\lambda \in F} \left\|(\lambda - L)^{-1}\right\| < \infty.$$

Exercise 2.9.5 Let L be a bounded linear operator on a Banach space V to V; and let $\lambda_0 \in \sigma(L)$ be an isolated nonzero eigenvalue of L. Let $\{L_n\}$ be a sequence of bounded linear operators on V to V with $\|L - L_n\| \to 0$ as $n \to \infty$. Assume $F \subset \rho(L)$ is closed and bounded. Prove that there exists N such that

$$n \geq N \quad \Longrightarrow \quad F \subset \rho(L_n).$$

This shows that approximating sequences $\{L_n\}$ cannot produce extraneous convergent sequences of approximating eigenvalues.
Hint: Use the result of Exercise 2.9.4.

Exercise 2.9.6 Assume L is a compact operator on V to V, a complex Banach space, and let $\{L_n\}$ be a sequence of approximating bounded linear compact operators with $\|L - L_n\| \to 0$ as $n \to \infty$. Referring to the curve $\Gamma = \{\lambda : |\lambda - \lambda_0| = \varepsilon\}$ of (2.9.5), we have from Exercise 2.9.5 that we can define

$$E(\sigma_n, L_n) = \frac{1}{2\pi i} \int_{|\lambda - \lambda_0| = \varepsilon} (\lambda - L_n)^{-1}\, d\lambda, \quad n \geq N$$

with σ_n denoting the portion of $\sigma(L_n)$ located within Γ. Prove

$$\|E(\sigma_n, L_n) - E(\lambda_0, L)\| \to 0 \quad \text{as } n \to \infty.$$

It can be shown that $\mathcal{R}(E(\sigma_n, L_n))$ consists of combinations of the ordinary and generalized eigenvectors of L_n corresponding to the eigenvalues of L_n within σ_n. In addition, prove that for every $u \in \mathcal{N}((\lambda - K)^{\nu(\lambda_0)})$,

$$E(\sigma_n, L_n)u \to u \quad \text{as } n \to \infty.$$

This shows convergence of approximating simple and generalized eigenfunctions of L_n to those of L.

Suggestion for Further Reading.

See "Suggestion for Further Readings" in Chapter 1.

3
Approximation Theory

In this chapter, we deal with the problem of approximation of functions. A prototype problem can be described as follows: For some function f, known exactly or approximately, find an approximation that has a more simply computable form, with the error of the approximation within a given error tolerance. Often the function f is not known exactly. For example, if the function comes from a physical experiment, we usually have a table of function values only. Even when a closed-form expression is available, it may happen that the expression is not easily computable, for example,

$$f(x) = \int_0^x e^{-t^2} dt.$$

The approximating functions need to be of simple form so that it is easy to make calculations with them. The most commonly used classes of approximating functions are the polynomials, piecewise polynomial functions, and trigonometric polynomials.

We begin with a review of some important theorems on the uniform approximation of continuous functions by polynomials. We then discuss several approaches to the construction of approximating functions. In Section 3.2, we define and analyze the use of interpolation functions. In Section 3.3 we discuss best approximation in general normed spaces, and in Section 3.4 we look at best approximation in inner product spaces. Section 3.5 is on the important special case of approximations using orthogonal polynomials, and Section 3.6 introduces approximations through projection operators. The chapter concludes with a discussion in Section 3.7 of

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_3,
© Springer Science + Business Media, LLC 2009

the uniform error bounds in polynomial and trigonometric approximations of smooth functions.

3.1 Approximation of continuous functions by polynomials

The classical Weierstrass Theorem is a fundamental result in the approximation of continuous functions by polynomials.

Theorem 3.1.1 (Weierstrass) *Let $f \in C[a,b]$. Then for any $\varepsilon > 0$, there exists a polynomial p for which*

$$\|f - p\|_\infty \leq \varepsilon.$$

The theorem states that any continuous function f can be approximated uniformly by polynomials, no matter how badly behaved f may be on $[a,b]$. Several proofs of this seminal result are given in [64, Chap. 6], including an interesting constructive result using Bernstein polynomials.

Various generalization of this classical result can be found in the literature. The following is one such general result. Its proof can be found in several textbooks on analysis or functional analysis (e.g., [44, pp. 420–422]).

Theorem 3.1.2 (Stone–Weierstrass) *Let $D \subset \mathbb{R}^d$ be a compact set. Suppose S is a subspace of $C(D)$, the space of continuous functions on D, with the following three properties.*

(a) *S contains all constant functions.*

(b) *$u, v \in S \Rightarrow uv \in S$.*

(c) *For each pair of points $\boldsymbol{x}, \boldsymbol{y} \in D$, $\boldsymbol{x} \neq \boldsymbol{y}$, there exists $v \in S$ such that $v(\boldsymbol{x}) \neq v(\boldsymbol{y})$.*

Then S is dense in $C(D)$, i.e., for any $v \in C(D)$, there is a sequence $\{v_n\} \subset S$ such that

$$\|v - v_n\|_{C(D)} \to 0 \quad \text{as } n \to \infty.$$

As simple consequences of Theorem 3.1.2, we have the next two results.

Corollary 3.1.3 *Let D be a compact set in $\mathbb{R}^d$. Then the set of all the polynomials on D is dense in $C(D)$.*

Corollary 3.1.4 *The set of all trigonometric polynomials is dense in the space $C_p([-\pi,\pi])$ of 2π-periodic continuous functions on $\mathbb{R}$.*

Obviously, Theorem 3.1.1 is a particular case of Corollary 3.1.3.

Denote by $\mathbb{P}([0,1])$ the space of all polynomials on $[0,1]$. Then Theorem 3.1.1 states that $\mathbb{P}([0,1])$ is dense in $C[0,1]$. This also implies that $\mathbb{P}([0,1])$ is dense in the space $L^2(0,1)$ since $C[0,1]$ is dense in $L^2(0,1)$ and $\|v\|_{L^2(0,1)} \le \|v\|_\infty$ for any $v \in C[0,1]$. An interesting question is whether proper subspaces of $\mathbb{P}([0,1])$ is also dense in $L^2(0,1)$. In this regard, the following result holds; for its proof, see e.g. [102, Section 6.2].

Theorem 3.1.5 (MUNTZ'S THEOREM) *Let $\lambda_1 < \lambda_2 < \cdots$ be a sequence of positive numbers with $\lambda_n \to \infty$ as $n \to \infty$. Then* $\operatorname{span}\{x^{\lambda_j}\}_{j\ge 1}$ *is dense in $L^2(0,1)$ if and only if $\sum_{j=1}^{\infty} \lambda_j^{-1} = \infty$.*

Note that in this theorem, the exponents $\{\lambda_j\}_{j\ge 1}$ are allowed to be non-integers.

Exercise 3.1.1 Prove Corollary 3.1.3 by applying Theorem 3.1.2.

Exercise 3.1.2 Prove Corollary 3.1.4 by applying Theorem 3.1.2.
Hint: $v \in C_p([-\pi,\pi])$ can be viewed as a continuous function on the unit circle in $\mathbb{R}^2$ when the argument t of v is identified with the point on the unit circle whose angle measured from x_1-axis is t.

Exercise 3.1.3 Prove Corollary 3.1.4 by applying Corollary 3.1.3.
Hint: Let D be the unit circle in $\mathbb{R}^2$, and consider the trigonometric polynomials as the restrictions to D of polynomials in two variables.

Exercise 3.1.4 Let D be a compact set in $\mathbb{R}^d$. Assume $f \in C(D)$ is such that $\int_D f(\boldsymbol{x})\, \boldsymbol{x}^\alpha dx = 0$ for any multi-index $\alpha = (\alpha_1, \cdots, \alpha_d)$. Then $f(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in D$.

Exercise 3.1.5 Show that every continuous function f defined on $[0,\infty)$ with the property $\lim_{x\to\infty} f(x) = 0$ can be approximated by a sequence of functions of the form

$$q_n(x) = \sum_{j=1}^{n} c_{n,j} e^{-jax},$$

where $a > 0$ is any fixed number, and $\{c_{n,j}\}$ are constants.
Hint: Apply Theorem 3.1.1 to the function

$$\begin{cases} f(-\log t/a), & 0 < t \le 1, \\ 0, & t = 0. \end{cases}$$

Exercise 3.1.6 Assume $f \in C([-1,1])$ is an even function. Show that $f(x)$ can be uniformly approximated on $[-1,1]$ by a sequence of polynomials of the form $p_n(x^2)$.
Hint: Consider $f(\sqrt{x})$ for $x \in [0,1]$.

Exercise 3.1.7 Let $f \in C^m[a,b]$, $m \geq 0$ integer. Show that there is a sequence of polynomials $\{p_n\}_{n\geq 1}$ such that

$$\|f - p_n\|_{C^m[a,b]} \to 0 \quad \text{as } n \to \infty.$$

Hint: Apply Theorem 3.1.1.

Exercise 3.1.8 Let $\Omega \subset \mathbb{R}^d$ be a domain, and $f \in C^m(\overline{\Omega})$, $m \geq 0$ integer. Show that there is a sequence of polynomials $\{p_n\}_{n\geq 1}$ such that

$$\|f - p_n\|_{C^m(\overline{\Omega})} \to 0 \quad \text{as } n \to \infty.$$

Hint: Apply Corollary 3.1.3.

3.2 Interpolation theory

We begin by discussing the interpolation problem in an abstract setting. Let V be a normed space over a field $\mathbb{K}$ of numbers ($\mathbb{R}$ or $\mathbb{C}$). Recall that the space of all the linear continuous functionals on V is called the *dual space* of V and is denoted by V' (see Section 2.5).

An abstract interpolation problem can be stated in the following form. Suppose V_n is an n-dimensional subspace of V, with a basis $\{v_1, \ldots, v_n\}$. Let $L_i \in V'$, $1 \leq i \leq n$, be n linear continuous functionals. Given n numbers $b_i \in \mathbb{K}$, $1 \leq i \leq n$, find $u_n \in V_n$ such that the interpolation conditions

$$L_i u_n = b_i, \quad 1 \leq i \leq n$$

are satisfied.

Some questions arise naturally: Does the interpolation problem have a solution? If so, is it unique? If the interpolation function is used to approximate a given function $f(x)$, what can be said about error in the approximation?

Definition 3.2.1 *We say that the functionals L_i, $1 \leq i \leq n$, are linearly independent over V_n if*

$$\sum_{i=1}^{n} a_i L_i(v) = 0 \quad \forall\, v \in V_n \quad \Longrightarrow \quad a_i = 0, \quad 1 \leq i \leq n.$$

Lemma 3.2.2 *The linear functionals $L_1, \ldots, L_n$ are linearly independent over V_n if and only if*

$$\det(L_i v_j)_{n\times n} = \det \begin{pmatrix} L_1 v_1 & \cdots & L_1 v_n \\ \vdots & \ddots & \vdots \\ L_n v_1 & \cdots & L_n v_n \end{pmatrix} \neq 0.$$

Proof. By definition,

$$\begin{aligned} & L_1, \ldots, L_n \text{ are linearly independent over } V_n \\ \iff & \sum_{i=1}^{n} a_i L_i(v_j) = 0,\ 1 \le j \le n \quad \Longrightarrow \quad a_i = 0,\ 1 \le i \le n \\ \iff & \det(L_i v_j) \neq 0. \end{aligned}$$

□

Theorem 3.2.3 *The following statements are equivalent:*

1. *The interpolation problem has a unique solution.*
2. *The functionals $L_1, \ldots, L_n$ are linearly independent over V_n.*
3. *The only element $u_n \in V_n$ satisfying*

$$L_i u_n = 0, \qquad 1 \le i \le n$$

is $u_n = 0$.

4. *For any data $\{b_i\}_{i=1}^n$, there exists one $u_n \in V_n$ such that*

$$L_i u_n = b_i, \quad 1 \le i \le n.$$

Proof. From linear algebra, for a square matrix $A \in \mathbb{K}^{n\times n}$, the following statements are equivalent:

1. The system $A\boldsymbol{x} = \boldsymbol{b}$ has a unique solution $\boldsymbol{x} \in \mathbb{K}^n$ for any $\boldsymbol{b} \in \mathbb{K}^n$.
2. $\det(A) \neq 0$.
3. If $A\boldsymbol{x} = \boldsymbol{0}$, then $\boldsymbol{x} = \boldsymbol{0}$.
4. For any $\boldsymbol{b} \in \mathbb{K}^n$, the system $A\boldsymbol{x} = \boldsymbol{b}$ has a solution $\boldsymbol{x} \in \mathbb{K}^n$.

The theorem now follow from these statements and Lemma 3.2.2. □

Now given $u \in V$, its interpolant $u_n = \sum_{i=1}^n a_i v_i$ in V_n is defined by the interpolation conditions

$$L_i u_n = L_i u, \quad 1 \le i \le n.$$

The coefficients $\{a_i\}_{i=1}^n$ can be found from the linear system

$$\begin{pmatrix} L_1 v_1 & \cdots & L_1 v_n \\ \vdots & \ddots & \vdots \\ L_n v_1 & \cdots & L_n v_n \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} L_1 u \\ \vdots \\ L_n u \end{pmatrix},$$

which has a unique solution if the functionals $L_1, \ldots, L_n$ are linearly independent over V_n.

An error analysis in the abstract framework is difficult to carry out. For a general discussion of such error analysis, see Davis [64, Chap. 3]. Here we only give error formulas or bounds for certain concrete situations.

3.2.1 Lagrange polynomial interpolation

Let f be a continuous function defined on a finite closed interval $[a, b]$. Let

$$\Delta : a \le x_0 < x_1 < \cdots < x_n \le b$$

be a partition of the interval $[a, b]$. Choose $V = C[a, b]$, the space of continuous functions $f : [a, b] \to \mathbb{K}$; and choose V_{n+1} to be $\mathbb{P}_n$, the space of the polynomials of degree less than or equal to n. Then the Lagrange interpolant of degree n of f is defined by the conditions

$$p_n(x_i) = f(x_i), \quad 0 \le i \le n, \quad p_n \in \mathbb{P}_n. \tag{3.2.1}$$

Here the interpolation linear functionals are

$$L_i f = f(x_i), \quad 0 \le i \le n.$$

If we choose the monomials $v_j(x) = x^j$, $0 \le j \le n$, as the basis for $\mathbb{P}_n$, then it can be shown that

$$\det(L_i v_j)_{(n+1)\times(n+1)} = \prod_{j>i}(x_j - x_i) \neq 0. \tag{3.2.2}$$

Thus there exists a unique Lagrange interpolation polynomial.

Furthermore, we have the representation formula

$$p_n(x) = \sum_{i=0}^{n} f(x_i)\,\phi_i(x), \quad \phi_i(x) \equiv \prod_{j\neq i} \frac{x - x_j}{x_i - x_j}, \tag{3.2.3}$$

called *Lagrange's formula* for the interpolation polynomial. The functions ϕ_i satisfy the special interpolation conditions

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

The functions $\{\phi_i\}_{i=0}^n$ form a basis for $\mathbb{P}_n$, and they are often called *Lagrange basis functions*. See Figure 3.1 for graphs of $\{\phi_i(x)\}_{i=0}^3$ for $n = 3$, the case of cubic interpolation, with even spacing on the interval $[1, 4]$.

Outside of the framework of Theorem 3.2.3, the formula (3.2.3) shows directly the existence of a solution to the Lagrange interpolation problem (3.2.1). The uniqueness result can also be proven by showing that the interpolant corresponding to the homogeneous data is zero. Let us show this. Let $p_n \in \mathbb{P}_n$ with $p_n(x_i) = 0$, $0 \le i \le n$. Then the polynomial p_n must contain the factors $(x - x_i)$, $1 \le i \le n$. Since $\deg(p_n) \le n$ and $\deg \Pi_{i=1}^n (x - x_i) = n$, we have

$$p_n(x) = c \prod_{i=1}^{n} (x - x_i)$$

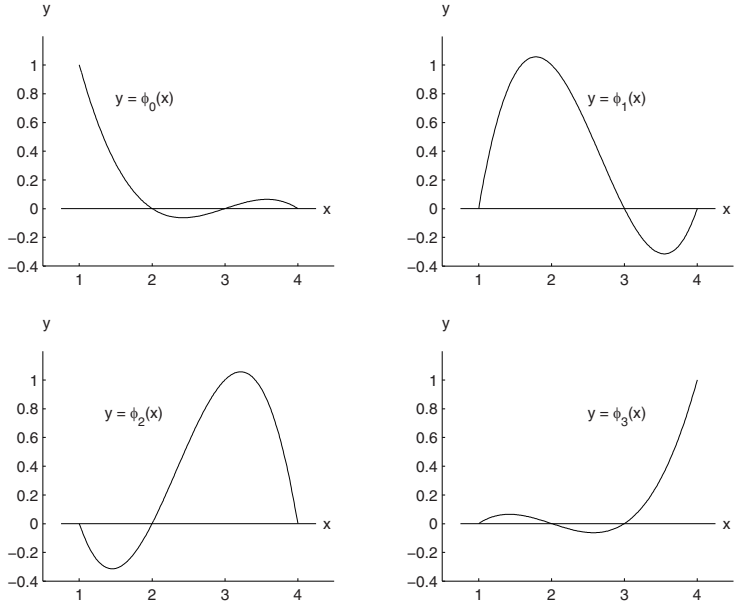

FIGURE 3.1. The Lagrange basis functions for $n = 3$, with nodes $\{1, 2, 3, 4\}$

for some constant c. Using the condition $p_n(x_0) = 0$, we see that $c = 0$ and therefore, $p_n \equiv 0$. We note that by Theorem 3.2.3, this result on the uniqueness of the solvability of the homogeneous problem also implies the existence of a solution.

In the above, we have indicated three methods for showing the existence and uniqueness of a solution to the interpolation problem (3.2.1). The method based on showing the determinant of the coefficient is nonzero, as in (3.2.2), can be done easily only in simple situations such as Lagrange polynomial interpolation. Usually it is simpler to show that the interpolant corresponding to the homogeneous data is zero, even for complicated interpolation conditions. For practical calculations, it is also useful to have a representation formula that is the analogue of (3.2.3), but such a formula is sometimes difficult to find.

These results on the existence and uniqueness of polynomial interpolation extend to the case that $\{x_0, \ldots, x_n\}$ are any $n + 1$ distinct points in the complex plane $\mathbb{C}$. The proofs remain the same.

For the interpolation error in Lagrange polynomial interpolation, we have the following result.

Proposition 3.2.4 *Assume $f \in C^{n+1}[a,b]$. Then there exists a point ξ_x between $\min_i\{x_i, x\}$ and $\max_i\{x_i, x\}$ such that*

$$f(x) - p_n(x) = \frac{\omega_n(x)}{(n+1)!} f^{(n+1)}(\xi_x), \quad \omega_n(x) = \prod_{i=0}^{n}(x - x_i). \tag{3.2.4}$$

Proof. The result is obvious if $x = x_i$, $0 \le i \le n$. Suppose $x \ne x_i$, $0 \le i \le n$, and denote

$$E(x) = f(x) - p_n(x).$$

Consider the function

$$g(t) = E(t) - \frac{\omega_n(t)}{\omega_n(x)} E(x).$$

We see that $g(t)$ has $(n+2)$ distinct roots, namely $t = x$ and $t = x_i$, $0 \le i \le n$. By the Mean Value Theorem, $g'(t)$ has $n+1$ distinct roots. Applying repeatedly the Mean Value Theorem to derivatives of g, we conclude that $g^{(n+1)}(t)$ has a root $\xi_x \in (\min_i\{x_i, x\}, \max_i\{x_i, x\})$. Then

$$0 = g^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - \frac{(n+1)!}{\omega_n(x)} E(x),$$

and the result is proved. □

There are other ways of looking at polynomial interpolation error. Using Newton divided differences, we can show

$$f(x) - p_n(x) = \omega_n(x)\, f[x_0, x_1, \ldots, x_n, x] \tag{3.2.5}$$

with $f[x_0, x_1, \ldots, x_n, x]$ a divided difference of f of order $n+1$. See [15, Section 3.2] for a development of this approach, together with a general discussion of divided differences and their use in interpolation.

We should note that high degree polynomial interpolation with a uniform mesh is likely to lead to problems. Figure 3.2 contains graphs of $\omega_n(x)$ for various degrees n. From these graphs, it is clear that the error behaviour is worse near the endpoint nodes than near the center node points. This leads to $p_n(x)$ failing to converge for such simple functions as $f(x) = (1+x^2)^{-1}$ on $[-5, 5]$, a famous example due to Carl Runge. A further discussion of this can be found in [15, Section 3.5]. In contrast, interpolation using the zeros of Chebyshev polynomials leads to excellent results. This is discussed in Section 3.7 of this chapter; and a further discussion is given in [15, p. 228].

3.2.2 Hermite polynomial interpolation

The main idea is to use values of both $f(x)$ and $f'(x)$ as interpolation conditions. Assume f is a continuously differentiable function on a finite

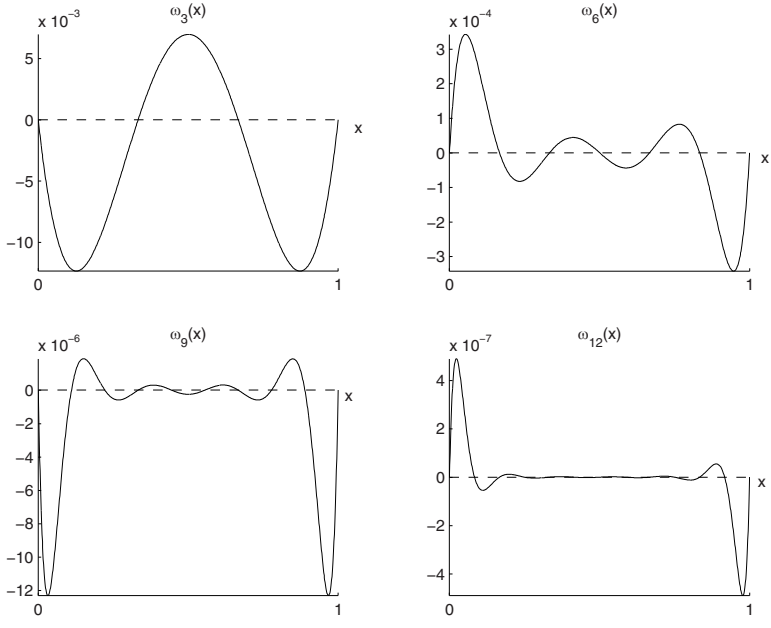

FIGURE 3.2. Examples of the polynomials $\omega_n(x)$ occurring in the interpolation error formulas (3.2.4) and (3.2.5)

interval $[a, b]$. Let

$$\Delta : a \le x_1 < \cdots < x_n \le b$$

be a partition of the interval $[a, b]$. Then the Hermite interpolant $p_{2n-1} \in \mathbb{P}_{2n-1}$ of degree less than or equal to $2n-1$ of f is chosen to satisfy

$$p_{2n-1}(x_i) = f(x_i), \quad p'_{2n-1}(x_i) = f'(x_i), \quad 1 \le i \le n. \tag{3.2.6}$$

We have results on Hermite interpolation similar to those for Lagrange interpolation, as given in Exercise 3.2.6.

More generally, for a given set of non-negative integers $\{m_i\}_{i=1}^n$, one can define a general Hermite interpolation problem as follows. Find $p_N \in \mathbb{P}_N(a, b)$, $N = \sum_{i=1}^n (m_i + 1) - 1$, to satisfy the interpolation conditions

$$p_N^{(j)}(x_i) = f^{(j)}(x_i), \quad 0 \le j \le m_i, \ 1 \le i \le n.$$

Again it can be shown that the interpolant with the homogeneous data is zero so that the interpolation problem has a unique solution. Also if $f \in C^{N+1}[a, b]$, then the error satisfies

$$f(x) - p_N(x) = \frac{1}{(N+1)!} \prod_{i=1}^{n} (x - x_i)^{m_i+1} f^{(N+1)}(\xi_x)$$

for some $\xi_x \in [a,b]$. For an illustration of an alternative error formula for the Hermite interpolation problem (3.2.6) that involves only the Newton divided difference of f, see [15, p. 161].

3.2.3 Piecewise polynomial interpolation

For simplicity, we focus our discussion on piecewise linear interpolation. Let $f \in C[a,b]$, and let

$$\Delta : a = x_0 < x_1 < \cdots < x_n = b$$

be a partition of the interval $[a,b]$. Denote $h_i = x_i - x_{i-1}$, $1 \le i \le n$, and $h = \max_{1\le i\le n} h_i$. The piecewise linear interpolant $\Pi_\Delta f$ of f is defined through the following two requirements:

- For $i = 1, \ldots, n$, $\Pi_\Delta f|_{[x_{i-1},x_i]}$ is linear.
- For $i = 0, 1, \ldots, n$, $\Pi_\Delta f(x_i) = f(x_i)$.

It is easy to see that $\Pi_\Delta f$ exists and is unique, and

$$\Pi_\Delta f(x) = \frac{x_i - x}{h_i} f(x_{i-1}) + \frac{x - x_{i-1}}{h_i} f(x_i), \quad x \in [x_{i-1}, x_i], \tag{3.2.7}$$

for $1 \le i \le n$.

For a general $f \in C[a,b]$, it is relatively straightforward to show

$$\max_{x\in[a,b]} |f(x) - \Pi_\Delta f(x)| \le \omega(f,h) \tag{3.2.8}$$

with $\omega(f,h)$ the *modulus of continuity* of f on $[a,b]$:

$$\omega(f,h) = \max_{\substack{|x-y|\le h \\ a\le x,y\le b}} |f(x) - f(y)|.$$

Suppose $f \in C^2[a,b]$. By using (3.2.4) and (3.2.7), it is straightforward to show that

$$\max_{x\in[a,b]} |f(x) - \Pi_\Delta f(x)| \le \frac{h^2}{8} \max_{x\in[a,b]} |f''(x)|. \tag{3.2.9}$$

Now instead of $f \in C^2[a,b]$, assume $f \in H^2(a,b)$ so that

$$\|f\|^2_{H^2(a,b)} = \int_a^b \left[|f(x)|^2 + |f'(x)|^2 + |f''(x)|^2\right] dx < \infty.$$

Here $H^2(a,b)$ is an example of Sobolev spaces. An introductory discussion was given in Examples 1.2.28 and 1.3.7. The space $H^2(a,b)$ consists of continuously differentiable functions f whose second derivative exists a.e. and belongs to $L^2(a,b)$. A detailed discussion of Sobolev spaces is given

in Chapter 7. We are interested in estimating the error in the piecewise linear interpolant $\Pi_\Delta f$ and its derivative $(\Pi_\Delta f)'$ under the assumption $f \in H^2(a,b)$.

We consider the error in the L^2 norm,

$$\begin{aligned}\|f - \Pi_\Delta f\|^2_{L^2(a,b)} &= \int_a^b |f(x) - \Pi_\Delta f(x)|^2 dx \\ &= \sum_{i=1}^n \int_{x_{i-1}}^{x_i} |f(x) - \Pi_\Delta f(x)|^2 dx.\end{aligned}$$

For a function $\widehat{f} \in H^2(0,1)$, let $\widehat{\Pi}\widehat{f}$ be its linear interpolant:

$$\widehat{\Pi}\widehat{f}(\xi) = \widehat{f}(0)\,(1-\xi) + \widehat{f}(1)\,\xi, \quad 0 \le \xi \le 1.$$

By Taylor's theorem,

$$\begin{aligned}\widehat{f}(0) &= \widehat{f}(\xi) - \xi\,\widehat{f}'(\xi) - \int_\xi^0 t\,\widehat{f}''(t)\,dt, \\ \widehat{f}(1) &= \widehat{f}(\xi) + (1-\xi)\,\widehat{f}'(\xi) + \int_\xi^1 (1-t)\,\widehat{f}''(t)\,dt.\end{aligned}$$

Thus

$$\widehat{f}(\xi) - \widehat{\Pi}\widehat{f}(\xi) = -\xi \int_\xi^1 (1-t)\,\widehat{f}''(t)\,dt - (1-\xi)\int_0^\xi t\,\widehat{f}''(t)\,dt,$$

and therefore

$$\int_0^1 |\widehat{f}(\xi) - \widehat{\Pi}\widehat{f}(\xi)|^2 d\xi \le c \int_0^1 |\widehat{f}''(\xi)|^2 d\xi \tag{3.2.10}$$

for some constant c independent of $\widehat{f}$. Using (3.2.10),

$$\begin{aligned}&\int_{x_{i-1}}^{x_i} |f(x) - \Pi_\Delta f(x)|^2 dx \\ &\quad = h_i \int_0^1 |f(x_{i-1} + h_i\xi) - \widehat{\Pi} f(x_{i-1} + h_i\xi)|^2 d\xi \\ &\quad \le c\,h_i \int_0^1 \left| \frac{d^2 f(x_{i-1} + h_i\xi)}{d\xi^2} \right|^2 d\xi \\ &\quad = c\,h_i^5 \int_0^1 |f''(x_{i-1} + h_i\xi)|^2 d\xi \\ &\quad = c\,h_i^4 \int_{x_{i-1}}^{x_i} |f''(x)|^2 dx.\end{aligned}$$

Therefore,

$$\|f - \Pi_\Delta f\|^2_{L^2(a,b)} = \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} |f(x) - \Pi_\Delta f(x)|^2 dx \le c\, h^4 \|f''\|^2_{L^2(a,b)},$$

i.e.

$$\|f - \Pi_\Delta f\|_{L^2(a,b)} \le c\, h^2 \|f''\|_{L^2(a,b)}. \tag{3.2.11}$$

A similar argument shows

$$\|f' - (\Pi_\Delta f)'\|_{L^2(a,b)} \le \widetilde{c}\, h\, \|f''\|_{L^2(a,b)}. \tag{3.2.12}$$

for another constant $\widetilde{c} > 0$.

In the theory of finite element interpolation, the above argument is generalized to error analysis of piecewise polynomial interpolation of any degree in higher spatial dimension.

3.2.4 Trigonometric interpolation

Another important and widely-used class of approximating functions are the trigonometric polynomials

$$p_n(x) = a_0 + \sum_{j=1}^{n} \left[a_j \cos(jx) + b_j \sin(jx)\right]. \tag{3.2.13}$$

If $|a_n| + |b_n| \neq 0$, we say $p_n(x)$ is a trigonometric polynomial of degree n. The function $p_n(x)$ is often considered as a function on the unit circle, in which case $p_n(\theta)$ would be a more sensible notation, with θ the central angle for a point on the unit circle. The set of the trigonometric polynomials of degree less than or equal to n is denoted by $\mathbb{T}_n$.

An equivalent way of writing such polynomials is as

$$p_n(x) = \sum_{j=-n}^{n} c_j e^{ijx}. \tag{3.2.14}$$

The equivalence is given by

$$a_0 = c_0, \quad a_j = c_j + c_{-j}, \quad b_j = i\,(c_j - c_{-j})\,.$$

Many computations with trigonometric polynomials are easier with (3.2.14) than with (3.2.13). With (3.2.14), we also can write

$$p_n(x) = \sum_{j=-n}^{n} c_j z^j = z^{-n} \sum_{k=0}^{2n} c_{k-n} z^k, \quad z = e^{ix}, \tag{3.2.15}$$

which brings us back to something involving polynomials.

The trigonometric polynomials of (3.2.13) are periodic with period 2π. Thus, we choose our interpolation nodes from the interval $[0, 2\pi)$ or any interval of length 2π:

$$0 \le x_0 < x_1 < \cdots < x_{2n} < 2\pi.$$

Often we use an even spacing, with

$$x_j = jh, \quad j = 0, 1, \ldots, 2n, \quad h = \frac{2\pi}{2n+1}. \tag{3.2.16}$$

The interpolation problem is to find a trigonometric polynomial $p_n(x)$ of degree less than or equal to n for which

$$p_n(x_j) = f_j, \quad j = 0, 1, \ldots, 2n \tag{3.2.17}$$

for given data values $\{f_j \mid 0 \le j \le 2n\}$. The existence and uniqueness of a solution of this problem can be reduced to that for Lagrange polynomial interpolation by means of the final formula in (3.2.15). Using it, we introduce the distinct complex nodes $z_j = e^{ix_j}$, $j = 0, 1, \ldots, 2n$. Then (3.2.17) can be rewritten as the polynomial interpolation problem

$$\sum_{k=0}^{2n} c_{k-n} z_j^k = z_j^n f_j, \quad j = 0, 1, \ldots, 2n.$$

All results from the polynomial interpolation problem with complex nodes can be applied to the trigonometric interpolation problem. For additional detail, see [15, Section 3.8]. Error bounds are given in Section 3.7 for the interpolation of a periodic function using trigonometric polynomials.

Example 3.2.5 Consider the periodic function

$$f(x) = e^{\sin x} \sin x. \tag{3.2.18}$$

Table 3.1 contains the maximum errors of the trigonometric interpolation polynomial $p_n(x)$ for varying values of n with the evenly spaced nodes defined in (3.2.16). □

Exercise 3.2.1 Show that there is a unique quadratic function p_2 satisfying the conditions

$$p_2(0) = a_0, \quad p_2(1) = a_1, \quad \int_0^1 p_2(x)\,dx = \overline{a}$$

with given a_0, a_1 and $\overline{a}$.

Exercise 3.2.2 Given a function f on $C[a, b]$, the moment problem is to find $p_n \in \mathbb{P}_n(a, b)$ such that

$$\int_a^b x^i p_n(x)\,dx = \int_a^b x^i f(x)\,dx, \quad 0 \le i \le n.$$

n	$\|f-p_n\|_\infty$	n	$\|f-p_n\|_\infty$
1	$1.16E+00$	8	$2.01E-07$
2	$2.99E-01$	9	$1.10E-08$
3	$4.62E-02$	10	$5.53E-10$
4	$5.67E-03$	11	$2.50E-11$
5	$5.57E-04$	12	$1.04E-12$
6	$4.57E-05$	13	$4.01E-14$
7	$3.24E-06$	14	$2.22E-15$

TABLE 3.1. Trigonometric interpolation errors for (3.2.18)

Show that the problem has a unique solution and the solution p_n satisfies

$$\int_a^b [f(x)-p_n(x)]^2\,dx \le \int_a^b [f(x)-q(x)]^2\,dx \quad \forall\, q \in \mathbb{P}_n(a,b).$$

Exercise 3.2.3 Let $x_0 < x_1 < x_2$ be three real numbers. Consider finding a polynomial $p(x)$ of degree ≤ 3 for which

$$\begin{aligned} p(x_0) &= y_0, \quad p(x_2) = y_2, \\ p'(x_1) &= y_1', \quad p''(x_1) = y_1'' \end{aligned}$$

with given data $\{y_0, y_2, y_1', y_1''\}$. Show there exists a unique such polynomial.

Exercise 3.2.4 Derive the formula (3.2.2) for the Vandermonde determinant of order $n+1$.
Hint: Introduce

$$V_n(x) = \det \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^n \\ 1 & x & x^2 & \cdots & x^n \end{pmatrix}.$$

Show

$$V_n(x) = V_{n-1}(x_{n-1})(x-x_0)\cdots(x-x_{n-1})$$

and use this to prove (3.2.2).

Exercise 3.2.5 Show that the Lagrange formula (3.2.3) can be rewritten in the form

$$p_n(x) = \frac{\displaystyle\sum_{j=0}^{n} \frac{w_j f(x_j)}{x-x_j}}{\displaystyle\sum_{j=0}^{n} \frac{w_j}{x-x_j}}$$

for x not a node point, for suitable values of $\{w_j\}$ which are dependent on only the nodes $\{x_j\}$. This formula is called the *barycentric representation* of $p_n(x)$.

Exercise 3.2.6 Show that the Hermite interpolation problem (3.2.6) admits a unique solution. Find a representation formula for the interpolant. Derive the error relation

$$f(x) - p_{2n-1}(x) = \frac{f^{(2n)}(\xi)}{(2n)!} \prod_{i=1}^{n} (x - x_i)^2$$

for some $\xi \in (\min_i\{x_i, x\}, \max_i\{x_i, x\})$, assuming $f \in C^{2n}[a, b]$.

Exercise 3.2.7 Show that if the modulus of continuity of f has the property $\omega(f, h) = o(h)$, i.e. $\omega(f, h)/h \to 0$ as $h \to 0+$, then f is a constant function.

Exercise 3.2.8 Let $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ be a partition of the interval $[a, b]$. Denote $x_{i-1/2} = (x_i + x_{i-1})/2$ and $h_i = x_i - x_{i-1}$ for $1 \le i \le n$, and $h = \max_{1 \le i \le n} h_i$. For $f \in C[a, b]$, its piecewise quadratic interpolant $\Pi_\Delta f$ is defined through two requirements:

(a) $\Pi_\Delta f|_{[x_{i-1}, x_i]}$ is quadratic for $i = 1, \ldots, n$;

(b) $\Pi_\Delta f(x_i) = f(x_i)$ for $i = 0, 1, \ldots, n$, and $\Pi_\Delta f(x_{i-1/2}) = f(x_{i-1/2})$ for $i = 1, 2, \ldots, n$.

Bound the error $\|f - \Pi_\Delta f\|_{L^2(a,b)}$ under the smoothness assumption $f^{(3)} \in L^2(a, b)$.

Exercise 3.2.9 Let us derive an error estimate for the composite trapezoidal rule, the convergence of which was discussed in Exercise 2.4.5. A standard error estimate is

$$|L_n v - Lv| \le c\, h^2 \max_{0 \le x \le 1} |v''(x)|$$

with $h = 1/n$. Assume $v'' \in L^1(0, 1)$. Use the idea in proving (3.2.11) to show that

$$|L_n v - Lv| \le c\, h^2 \|v''\|_{L^1(0,1)},$$

i.e. the smoothness requirement on the integrand can be weakened while the same order error estimate is kept. Improved estimates of this kind are valid for errors of more general numerical quadratures.

Exercise 3.2.10 An elementary argument for the improved error estimate of Exercise 3.2.9 is possible, under additional smoothness assumption on the integrand. Suppose $v \in C^2[a, b]$. For the composite trapezoidal rule, show that

$$L_n v - Lv = \int_a^b K_T(x)\, v''(x)\, dx,$$

where the Peano kernel function K_T is defined by

$$K_T(x) = \frac{1}{2}\,(x - x_{k-1})\,(x_k - x), \quad x_{k-1} \le x \le x_k$$

for $k = 1, 2, \ldots, n$. Use this relation to prove the quadrature error bound

$$|L_n v - Lv| \le c\, h^2 \|v''\|_{L^1(a,b)}.$$

Exercise 3.2.11 As another example of similar nature, show that for the composite Simpson's rule, the following error representation is valid:

$$L_n v - Lv = \int_a^b K_S(x)\, v^{(4)}(x)\, dx,$$

where the Peano kernel function K_S is defined by

$$K_S(x) = \begin{cases} h\,(x - x_{k-2})^3/18 - (x - x_{k-2})^4/24, & x_{k-2} \le x \le x_{k-1}, \\ h\,(x_k - x)^3/18 - (x_k - x)^4/24, & x_{k-1} \le x \le x_k \end{cases}$$

for $k = 2, 4, \ldots, n$. Use this relation to prove the quadrature error bound

$$|L_n v - Lv| \le c\, h^4 \|v^{(4)}\|_{L^1(0,1)}.$$

Exercise 3.2.12 Consider the computation of the integral

$$I = \int_a^b f(x)\, dx, \quad f \in C^1[a, b].$$

Divide the interval $[a, b]$ into n equal parts, and denote $h = (b-a)/n$ the meshsize and $x_k = a + k\,h$, $k = 0, \ldots, n$, the nodes. Let p be the piecewise cubic Hermite interpolant of $f(x)$, i.e., for any $k = 1, \cdots, n$, $p|_{[x_{k-1}, x_k]}$ is a polynomial of degree less than or equal to 3, and p satisfies the following interpolation conditions

$$p(x_k) = f(x_k), \quad p'(x_k) = f'(x_k), \quad k = 0, \ldots, n.$$

Then we approximate the integral by

$$\int_a^b f(x)\, dx \approx I_n = \int_a^b p(x)\, dx.$$

Show that

$$I_n = h\left[\frac{1}{2} f(a) + \sum_{k=1}^{n-1} f(x_k) + \frac{1}{2} f(b)\right] - \frac{h^2}{12}\left[f'(b) - f'(a)\right].$$

This formula is called the *corrected trapezoidal rule.*

Derive estimates for the error $|I - I_n|$ under the assumptions $f^{(4)} \in L^1(a, b)$ and $f^{(3)} \in L^1(a, b)$, respectively.

Exercise 3.2.13 (a) For the nodes $\{x_j\}$ of (3.2.16), show the identity

$$\sum_{j=0}^{2n} e^{ikx_j} = \begin{cases} 2n+1, & e^{ix_k} = 1, \\ 0, & e^{ix_k} \ne 1 \end{cases}$$

for $k \in \mathbb{Z}$ and $x_k = 2\pi k/(2n+1)$.
(b) Find the trigonometric interpolation polynomial that solves the problem (3.2.17) with the evenly spaced nodes of (3.2.16). Consider finding the interpolation polynomial in the form of (3.2.14). Show that the coefficients $\{c_j\}$ are given by

$$c_\ell = \frac{1}{2n+1} \sum_{j=0}^{2n} b_j e^{-i\ell x_j}, \quad \ell = -n, \ldots, n.$$

Hint: Use the identity in part (a) to solve the linear system

$$\sum_{k=-n}^{n} c_k e^{ikx_j} = b_j, \quad j = 0, 1, \ldots, 2n.$$

Begin by multiplying equation j by $e^{-i\ell x_j}$, and then sum the equations over j.

Exercise 3.2.14 Prove the error bound (3.2.8).

Exercise 3.2.15 Let $\{z_k\}_{k=0}^{2n}$ be distinct non-zero complex numbers. Show that for any given $(2n+1)$ complex numbers $\{y_k\}_{k=0}^{2n}$, we can uniquely determine the coefficients $\{c_j\}_{j=-n}^{n}$ of the function

$$f(z) = \sum_{j=-n}^{n} c_j z^j$$

from the interpolation conditions

$$f(z_k) = y_k, \quad 0 \le k \le 2n.$$

Exercise 3.2.16 Continuing Exercise 3.2.15, suppose $z_k = e^{ix_k}$, $0 \le x_0 < x_1 < \cdots < x_{2n} < 2\pi$. If $\{y_k\}_{k=0}^{2n}$ are real, show that

$$f(e^{ix}) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \sin(kx)]$$

for real coefficients $\{a_k\}_{k=0}^{n}$ and $\{b_k\}_{k=1}^{n}$.

3.3 Best approximation

We approximate a function $f(x)$ by choosing some member from a restricted class of functions. For example, a polynomial was selected from $\mathbb{P}_n$ by using interpolation to $f(x)$. It is useful to consider the best that can be done with such a class of approximating functions: How small an error is possible when selecting an approximation from the given class of approximating functions? This is known as the *best approximation problem*. The solution depends on the function f, on the class of approximating functions, and on the norm by which the error is being measured. The best known cases use the uniform norm $\|\cdot\|_\infty$, the L^1-norm, and the L^2-norm (and other Hilbert space norms). We examine the best approximation problem in this section and some of the following sections.

Throughout this section, V is allowed to be either a real or complex linear space.

3.3.1 Convexity, lower semicontinuity

A best approximation problem can be described by the minimization of a certain functional measuring the error size, and some rather general results can be given within such a framework. We begin by introducing some useful concepts.

Definition 3.3.1 *Let V be a real or complex linear space, $K \subset V$. The set K is said to be* convex *if*

$$u, v \in K \quad \Longrightarrow \quad \lambda\, u + (1-\lambda)\, v \in K \quad \forall\, \lambda \in (0,1).$$

Informally, convexity of the set K is characterized by the property that the line segment joining any two elements of K is also contained in K.

If K is convex, by induction we can show

$$u_i \in K, \quad 1 \le i \le n \quad \Longrightarrow \quad \sum_{i=1}^{n} \lambda_i u_i \in K \quad \forall\, \lambda_i \ge 0 \text{ with } \sum_{i=1}^{n} \lambda_i = 1. \tag{3.3.1}$$

The expression $\sum_{i=1}^{n} \lambda_i u_i$ with $\lambda_i \ge 0$ and $\sum_{i=1}^{n} \lambda_i = 1$, is called a *convex combination* of $\{u_i\}_{i=1}^{n}$.

Definition 3.3.2 *Let K be a convex set in a linear space V. A function $f : K \to \mathbb{R}$ is said to be* convex *if*

$$f(\lambda\, u + (1-\lambda)\, v) \le \lambda\, f(u) + (1-\lambda)\, f(v) \quad \forall\, u, v \in K,\ \forall\, \lambda \in [0,1].$$

The function f is strictly convex *if the above inequality is strict for $u \ne v$ and $\lambda \in (0,1)$.*

To obtain a more intuitive sense of what it means for a function f to be convex, interpret it geometrically for the graph of a real-valued convex function f over $\mathbb{R}^2$. If any two points u and v in $\mathbb{R}^2$ are connected by a straight line segment L, then any point on the line segment joining $(u, f(u))$ and $(v, f(v))$, denoted as $(\lambda\, u + (1-\lambda)\, v, \lambda\, f(u) + (1-\lambda)\, f(v))$ for some $\lambda \in [0,1]$, is located above the corresponding point $(\lambda\, u + (1-\lambda)\, v, f(\lambda\, u + (1-\lambda)\, v))$ on the graph of f. The reader should note that the phrase "strictly convex" has another meaning in the literature on approximation theory, related somewhat to our definition but still distinct from it.

Definition 3.3.3 *Let V be a normed space. A set $K \subset V$ is* closed *if $\{v_n\} \subset K$ and $v_n \to v$ imply $v \in K$. The set K is* weakly closed *if $\{v_n\} \subset K$ and $v_n \rightharpoonup v$ imply $v \in K$.*

A weakly closed set is certainly closed. But the converse statement is not true on general infinite dimensional spaces.

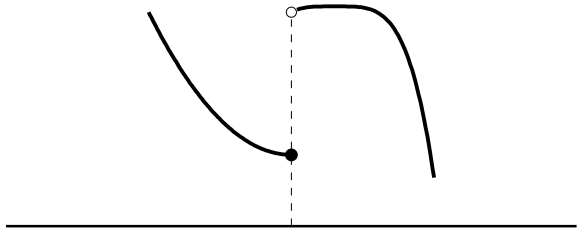

FIGURE 3.3. A l.s.c. function with discontinuity

Definition 3.3.4 *Let V be a normed space, $K \subset V$. A function $f : K \to \mathbb{R}$ is* (sequentially) lower semicontinuous (l.s.c.) *if $\{v_n\} \subset K$ and $v_n \to v \in K$ imply*

$$f(v) \leq \liminf_{n\to\infty} f(v_n).$$

The function f is weakly sequentially lower semicontinuous *or* weakly lower semicontinuous (w.l.s.c.) *if the above inequality is valid for any sequence $\{v_n\} \subset K$ with $v_n \rightharpoonup v \in K$.*

Obviously continuity implies lower semicontinuity. The converse statement is not true, as lower semicontinuity allows discontinuity in a function; see Figure 3.3 for such an example. It is easily seen that if f is w.l.s.c., then it is l.s.c. The notion of weak lower semicontinuity is very useful in a number of topics with applied and computational mathematics, including the study of boundary value problems for elliptic partial differential equations.

Example 3.3.5 We examine an example of a w.l.s.c. function. Let V be a normed space and let us show that the norm function is w.l.s.c. For this, let $\{v_n\} \subset V$ be a weakly convergent sequence, $v_n \rightharpoonup v \in V$. By Corollary 2.5.6, there is an $\ell \in V'$ such that $\ell(v) = \|v\|$ and $\|\ell\| = 1$. We notice that

$$\ell(v_n) \leq \|\ell\| \, \|v_n\| = \|v_n\|.$$

Therefore,

$$\|v\| = \ell(v) = \lim_{n\to\infty} \ell(v_n) \leq \liminf_{n\to\infty} \|v_n\|.$$

So $\|\cdot\|$ is w.l.s.c. (cf. Exercise 2.7.2).

In an inner product space, a simpler proof is possible to show the norm function is w.l.s.c. Indeed, assume V is an inner product space, and let $\{v_n\} \subset V$ be a weakly convergent sequence, $v_n \rightharpoonup v$. Then

$$\|v\|^2 = (v, v) = \lim_{n\to\infty} (v, v_n) \leq \liminf_{n\to\infty} \|v\| \, \|v_n\|,$$

and we easily obtain

$$\|v\| \leq \liminf_{n\to\infty} \|v_n\|.$$

See Exercise 3.3.6 for yet another proof. □

We now present a useful result on geometric functional analysis derived from the generalized Hahn-Banach theorem, concerning separation of convex sets.

Definition 3.3.6 *Let V be a real normed space, and A and B be non-empty sets in V. The sets A and B are said to be* separated *if there is a non-zero linear continuous functional ℓ on V and a number $\alpha \in \mathbb{R}$ such that*

$$\begin{aligned} \ell(u) &\leq \alpha \quad \forall\, u \in A, \\ \ell(v) &\geq \alpha \quad \forall\, v \in B. \end{aligned}$$

If the inequalities are strict, then we say the sets A and B are strictly separated.

The next result follows from Theorem 2.5.5. A proof of the result can be found in, e.g., [75].

Theorem 3.3.7 *Let V be a real normed space, A and B be two non-empty disjoint convex subsets of V such that one of them is compact, and the other is closed. Then the sets A and B can be strictly separated.*

This result is used later in Section 11.3.

3.3.2 Some abstract existence results

Given a real space V, a subset $K \subset V$, and a functional $f : K \to \mathbb{R}$, we consider the problem of finding a minimizer $v = u$ for $f(v)$ over K:

$$\inf_{v \in K} f(v). \tag{3.3.2}$$

A general reference for the results of this subsection is [247], including proofs of most of the results given here.

Before we present a general result on the existence of a solution to the problem (3.3.2), let us recall a classical result of Weierstrass: A real-valued continuous function f on a bounded closed interval $[a, b]$ $(-\infty < a < b < \infty)$ has a maximum and a minimum. We recall the main steps in a typical proof of the result for the part regarding a minimum. In the first step, we denote $K = [a, b]$ and

$$\alpha = \inf_{x \in K} f(x).$$

Then by the definition of infimum, there is a sequence $\{x_n\} \subset K$ such that $f(x_n) \to \alpha$ as $n \to \infty$. In the second step, we notice that the bounded closed interval K is compact. Thus, there is a subsequence $\{x_{n'}\} \subset \{x_n\}$ and some $x_0 \in [a, b]$ such that

$$x_{n'} \to x_0 \quad \text{as } n' \to \infty.$$

In the last step, we use the continuity of the function f:

$$f(x_0) = \lim_{n' \to \infty} f(x_{n'}) = \alpha,$$

i.e., x_0 is a minimizer of f on $K = [a, b]$.

Now we try to extend the above argument to the problem (3.3.2) on a general setting. The first step depends only on the definition of infimum. For the second step, we note that in an infinite-dimensional Banach space V, a bounded sequence does not necessarily contain a convergent subsequence (see Example 2.7.3). Nevertheless, if V is a reflexive Banach space, then Theorem 2.7.5 states that a bounded sequence in V contains a weakly convergent subsequence. Therefore, for the problem (3.3.2), we assume V is reflexive, K is bounded and weakly closed. This last condition ensures that the weak limit of a weakly convergent subsequence in K lies in K. Since the candidate of a minimizer is now only a weak limit, in the last step, we would assume the continuity of the functional f with respect to a weak limit. This assumption on f is too strong and too restrictive. Nevertheless, we notice that it is sufficient to assume f is weakly sequentially l.s.c. This assumption allows f to be discontinuous.

With the above consideration, we see that the conditions of the next result are quite natural.

Theorem 3.3.8 *Assume V is a reflexive Banach space, and assume $K \subset V$ is bounded and weakly closed. If $f : K \to \mathbb{R}$ is weakly sequentially l.s.c., then the problem* (3.3.2) *has a solution.*

Proof. Denote

$$\alpha = \inf_{v \in K} f(v).$$

By the definition of infimum, there exists a sequence $\{u_n\} \subset K$ with

$$f(u_n) \to \alpha \quad \text{as } n \to \infty.$$

The sequence $\{u_n\}$ is usually called a *minimizing* sequence of f over K. Since K is bounded, $\{u_n\}$ is a bounded sequence in the space V. Now V is reflexive, Theorem 2.7.5 implies that there exists a subsequence $\{u_{n'}\} \subset \{u_n\}$ which converges weakly to $u \in V$. Since K is weakly closed, we have $u \in K$; and since f is weakly sequentially l.s.c., we have

$$f(u) \le \liminf_{n' \to \infty} f(u_{n'}).$$

Therefore, $f(u) = \alpha$, and u is a solution of the minimization problem (3.3.2). Note that this proof also shows α is finite, $\alpha > -\infty$. □

In the above theorem, K is assumed to be bounded. Often we have the situation where K is unbounded (a subspace, for example). We can drop the boundedness assumption on K, and as a compensation we assume f to be coercive over K.

Definition 3.3.9 *Let V be a normed space, $K \subset V$. A real-valued functional f on V is said to be* coercive *over K if*

$$f(v) \to \infty \quad \text{as } \|v\| \to \infty,\ v \in K.$$

Theorem 3.3.10 *Assume V is a reflexive Banach space, $K \subset V$ is weakly closed. If $f : K \to \mathbb{R}$ is weakly sequentially l.s.c. and coercive on K, then the problem* (3.3.2) *has a solution.*

Proof. Pick any $v_0 \in K$ and define

$$K_0 = \{v \in K \mid f(v) \le f(v_0)\}.$$

Since f is coercive, K_0 is bounded. Since K is weakly closed and f is weakly sequentially l.s.c., we see that K_0 is weakly closed. The problem (3.3.2) is equivalent to

$$\inf_{v \in K_0} f(v)$$

which has at least one solution from Theorem 3.3.8. □

These results are rather general in nature. In applications, it is usually not convenient to verify the conditions associated with weakly convergent sequences. We replace these conditions by ones easier to verify. First we record a result of fundamental importance in convex analysis. A proof of the result is given in [76, p. 6].

Theorem 3.3.11 (Mazur Lemma) *Assume V is a normed space, and assume $\{v_n\}_{n\ge1}$ is a sequence converging weakly to u. Then there is a sequence $\{u_n\}_{n\ge1}$ of convex combinations of $\{v_n\}_{n\ge1}$,*

$$u_n = \sum_{i=n}^{N(n)} \lambda_i^{(n)} v_i, \quad \sum_{i=n}^{N(n)} \lambda_i^{(n)} = 1,\ \lambda_i^{(n)} \ge 0,\ n \le i \le N(n),$$

which converges strongly to u.

It is left as Exercise 3.3.6 to prove the following corollaries of Mazur Lemma.

- If K is convex and closed, then it is weakly closed.
- If f is convex and l.s.c. (or continuous), then it is weakly sequentially l.s.c.

Now we have the following variants of the existence results, and they are sufficient for our applications.

Theorem 3.3.12 *Assume V is a reflexive Banach space, $K \subset V$ is convex and closed, and $f : K \to \mathbb{R}$ is convex and l.s.c. If either*
(a) *K is bounded*
or
(b) *f is coercive on K,*
then the minimization problem (3.3.2) *has a solution. Furthermore, if f is strictly convex, then a solution to the problem* (3.3.2) *is unique.*

Proof. It remains to show that if f is strictly convex, then a minimizer of f over K is unique. Let us argue by contradiction. Assume there were two minimizers $u_1 \neq u_2$, with $f(u_1) = f(u_2)$ the minimal value of f on K. Since K is convex, $(u_1 + u_2)/2 \in K$. By the strict convexity of f, we would have

$$f\left(\frac{u_1 + u_2}{2}\right) < \frac{1}{2}\left[f(u_1) + f(u_2)\right] = f(u_1).$$

This relation contradicts the assumption that u_1 is a minimizer. □

In certain applications, the space V is not reflexive (e.g., $V = C[a, b]$). In such a case the above theorems are not applicable. Nevertheless, we notice that the reflexivity of V is used only to extract a weakly convergent subsequence from a bounded sequence in K. Also notice that we only need the completeness of the subset K, not that of the space V. Hence, we may modify Theorem 3.3.12 as follows.

Theorem 3.3.13 *Assume V is a normed space, $K \subset V$ is a convex and closed finite-dimensional subset, and $f : K \to \mathbb{R}$ is convex and l.s.c. If either*
(a) *K is bounded*
or
(b) *f is coercive on K,*
then the minimization problem (3.3.2) *has a solution. Furthermore, if f is strictly convex, then a solution to the problem* (3.3.2) *is unique.*

A subset is said to be finite-dimensional if it is a subset of a finite-dimensional subspace.

3.3.3 Existence of best approximation

Let us apply the above results to a best approximation problem. Let $u \in V$. We are interested in finding elements from $K \subset V$ which are closest to u among the elements in K. More precisely, we are interested in the minimization problem

$$\inf_{v \in K} \|u - v\|. \tag{3.3.3}$$

Obviously (3.3.3) is a problem of the form (3.3.2) with

$$f(v) = \|u - v\|.$$

Certainly $f(v)$ is convex and continuous (and hence l.s.c.). Furthermore, $f(v)$ is coercive if K is unbounded. We thus have the following existence theorems on best approximations.

Theorem 3.3.14 *Assume $K \subset V$ is a closed, convex subset of a reflexive Banach space V. Then there is an element $\widehat{u} \in K$ such that*

$$\|u - \widehat{u}\| = \inf_{v \in K} \|u - v\|.$$

Theorem 3.3.15 *Assume $K \subset V$ is a convex and closed finite-dimensional subset of a normed space V. Then there is an element $\widehat{u} \in K$ such that*

$$\|u - \widehat{u}\| = \inf_{v \in K} \|u - v\|.$$

In particular, a finite-dimensional subspace is both convex and closed.

Theorem 3.3.16 *Assume K is a finite-dimensional subspace of the normed space V. Then there is an element $\widehat{u} \in K$ such that*

$$\|u - \widehat{u}\| = \inf_{v \in K} \|u - v\|.$$

Example 3.3.17 Let $V = C[a, b]$ (or $L^p(a, b)$) and $K = \mathbb{P}_n$, the space of all the polynomials of degree less than or equal to n. Associated with the space V, we may use $L^p(a, b)$ norms, $1 \leq p \leq \infty$. The previous results ensure that for any $f \in C[a, b]$ (or $L^p(a, b)$), there exists a polynomial $f_n \in \mathbb{P}_n$ such that

$$\|f - f_n\|_{L^p(a,b)} = \inf_{q_n \in \mathbb{P}_n} \|f - q_n\|_{L^p(a,b)}.$$

Certainly, for different value of p, we have a different best approximation f_n. When $p = \infty$, f_n is called a "best uniform approximation of f". □

The existence of a best approximation from a finite dimensional subspace can also be proven directly. To do so, reformulate the minimization problem as a problem of minimizing a non-negative continuous real-valued function over a closed bounded subset of $\mathbb{R}^n$ or $\mathbb{C}^n$, and then appeal to the Heine-Borel Theorem from elementary analysis (see Theorem 1.6.2). This is left as Exercise 3.3.7.

3.3.4 Uniqueness of best approximation

Showing uniqueness requires greater attention to the properties of the norm or to the characteristics of the approximating subset K.

Arguing as in the proof for the uniqueness part in Theorem 3.3.12, we can easily show the next result.

Theorem 3.3.18 *Assume V is a normed space, and further assume that the function $f(v) = \|v\|^p$ is strictly convex for some $p \geq 1$. Let K be a convex subset of V. Then for any $u \in V$, a best approximation $\widehat{u}$ from K is unique.*

If V is an inner product space, then $f(v) = \|v\|^2$ is a strictly convex function on V (see Exercise 3.3.8), and therefore a solution to the best approximation problem in an inner product space is unique (provided it exists).

For $p \in (1, \infty)$, the function $f(v) = \|v\|^p_{L^p(\Omega)}$ is strictly convex on the space $L^p(\Omega)$. This is deduced from the Clarkson inequalities (1.5.5) and (1.5.6). Therefore, in $L^p(\Omega)$, $1 < p < \infty$, there can be at most one best approximation.

Notice that the strict convexity of the norm is a sufficient condition for the uniqueness of a best approximation, but the condition is not necessary. For example, the norm $\|\cdot\|_{L^\infty(a,b)}$ is not strictly convex, yet there are classical results stating that a best uniform approximation is unique for important classes of approximating functions. The following is the best known of such results.

Theorem 3.3.19 (Chebyshev Equi-oscillation Theorem) *Let $f \in C[a, b]$ for a finite interval $[a, b]$, and let $n \geq 0$ be an integer. Then there is a unique solution $\widehat{p}_n \in \mathbb{P}_n$ to the minimization problem*

$$\rho_n(f) = \min_{p \in \mathbb{P}_n} \|f - p\|_\infty .$$

It is characterized uniquely as follows. There is a set of $n + 2$ numbers

$$a \leq x_0 < x_1 < \cdots < x_{n+1} \leq b,$$

not necessarily unique, for which

$$f(x_j) - \widehat{p}_n(x_j) = \sigma\,(-1)^j \rho_n(f), \quad j = 0, 1, \ldots, n + 1$$

with $\sigma = +1$ or -1.

Theorem 3.3.20 *Let g be a continuous 2π-periodic function on $\mathbb{R}$, and let $n \geq 0$ be an integer. Then there is a unique trigonometric polynomial $\widehat{q}_n \in \mathbb{T}_n$ of degree less than or equal to n (see* (3.2.13)*) satisfying the minimization property*

$$\rho_n(g) = \min_{q \in \mathbb{T}_n} \|g - q\|_\infty.$$

Proofs of these two theorems are given in Meinardus [169, Section 3] and Davis [64, Chap. 7]. We return to these best uniform approximations in Section 3.7, where we look at the size of $\rho_n(f)$ as a function of the smoothness of f.

Another approach to proving the uniqueness of a best approximation is through the notion of a strictly normed space ([153]). A space V is said to be strictly normed if

$$\|u+v\| = \|u\| + \|v\| \quad \text{and} \quad u \neq 0$$

implies $v = \lambda u$ for some non-negative scalar λ.

Theorem 3.3.21 *Assume V is a strictly normed space, and $K \subset V$ is non-empty and convex. Then for any $u \in V$, there is at most one element $u_0 \in K$ such that*

$$\|u - u_0\| = \inf_{v \in K} \|u - v\|.$$

Proof. Denote $d = \inf_{v\in K} \|u - v\|$. Suppose there is another element $u_1 \in K$ with the property

$$\|u - u_1\| = d.$$

Since K is convex, $(u_0 + u_1)/2 \in K$ and so

$$\|(u - u_0)/2 + (u - u_1)/2\| = \|u - (u_0 + u_1)/2\| \geq d.$$

On the other hand,

$$\|(u - u_0)/2 + (u - u_1)/2\| \leq \|u - u_0\|/2 + \|u - u_1\|/2 = d.$$

Hence,

$$\|(u - u_0)/2 + (u - u_1)/2\| = \|(u - u_0)/2\| + \|(u - u_1)/2\|.$$

If $u \in K$, then obviously $u_1 = u_0 = u$. Otherwise, $u - u_0 \neq 0$ and since V is strictly normed, there is a $\lambda \geq 0$ with

$$\frac{1}{2}(u - u_1) = \lambda \frac{1}{2}(u - u_0).$$

From this, $\|u - u_1\| = \lambda \|u - u_0\|$, i.e., $d = \lambda d$. So $\lambda = 1$, and $u_1 = u_0$. □

It can be shown that an inner product space is strictly normed (see Exercise 3.3.9), and for $p \in (1, \infty)$, the space $L^p(\Omega)$ is also strictly normed. Thus, we can again conclude the uniqueness of a best approximation both in an inner product space and in $L^p(a, b)$ for $1 < p < \infty$.

Exercise 3.3.1 Prove (3.3.1).

Exercise 3.3.2 Let K be a set in a linear space V. Define a set

$$K_c = \left\{ \sum_{i=1}^{n} \lambda_i v_i \;\middle|\; v_i \in K,\ \lambda_i \geq 0,\ \sum_{i=1}^{n} \lambda_i = 1,\ n = 1, 2, \cdots \right\}.$$

Show that $K \subset K_c$, K_c is convex and is the smallest convex set containing K.

The set K_c is called the convex hull of K.

Exercise 3.3.3 Show that the intersection of any collection of convex sets, the sum of two convex sets, and a scalar multiple of a convex set are all convex.

Recall that for two sets V_1, V_2 in a linear space V, their sum is

$$V_1 + V_2 = \{v_1 + v_2 \mid v_1 \in V_1,\ v_2 \in V_2\}.$$

For a scalar α and a set K in the linear space V, the scalar multiple of the set is

$$\alpha\, K = \{\alpha\, v \mid v \in K\}.$$

Exercise 3.3.4 In certain context, it is convenient to consider convex functions defined on a whole linear space V, and for this purpose, the functions are allowed to take on the value $+\infty$. A function $F : V \to \mathbb{R} \cup \{+\infty\}$ is said to be convex if

$$F(\lambda\, u + (1-\lambda)\, v) \le \lambda\, F(u) + (1-\lambda)\, F(v) \quad \forall\, u, v \in V,\ \forall\, \lambda \in [0,1].$$

Let $K \subset V$, and $f : K \to \mathbb{R}$ be given. Define

$$F(v) = \begin{cases} f(v) & \text{if } v \in K, \\ +\infty & \text{otherwise.} \end{cases}$$

Show that $F : V \to \mathbb{R} \cup \{+\infty\}$ is convex if and only if K is a convex set in V and $f : K \to \mathbb{R}$ is a convex function.

Exercise 3.3.5 Let $g \in C[0,1]$ and let $n \ge 0$ be an integer. Define

$$E(g) \equiv \inf_{\deg(p) \le n} \left[\max_{0 \le x \le 1} \left(1 + x^2\right) |g(x) - p(x)| \right]$$

with $p(x)$ denoting a polynomial. Consider the minimization problem of finding at least one polynomial $\widehat{p}(x)$ of degree at most n for which

$$E(g) = \max_{0 \le x \le 1} \left(1 + x^2\right) |g(x) - \widehat{p}(x)|\,.$$

What can you say about the solvability of this problem?

Exercise 3.3.6 Apply Mazur Lemma to show that in a normed space, a convex closed set is weakly closed, and a convex l.s.c. function is w.l.s.c. In particular, if a sequence $\{v_n\}$ in the normed space converges weakly to an element v, then

$$\|v\| \le \liminf_{n \to \infty} \|v_n\|.$$

Exercise 3.3.7 Give a direct proof of Theorem 3.3.16, as discussed following Example 3.3.17.

Exercise 3.3.8 Show that in an inner product space V, the function $f(v) = \|v\|^2$ is strictly convex.

Exercise 3.3.9 Show that an inner product space is strictly normed.
Hint: Square and expand both sides of $\|u + v\| = \|u\| + \|v\|$. Then consider the implication of an equality in the Schwarz inequality.

3.4 Best approximations in inner product spaces, projection on closed convex sets

In an inner product space V, the norm $\|\cdot\|$ is induced by an associated inner product. From the discussions in the previous section, Theorem 3.3.18 and Exercise 3.3.8, or Theorem 3.3.21 and Exercise 3.3.9, a best approximation is unique. Alternatively, the uniqueness of the best approximation can be verified using the following characterization (3.4.1) of a best approximation when the norm is induced by an inner product.

Throughout this section, we assume V is a real inner product space. Many of the results generalize, and in some cases, they are stated in a more general form. A general reference for the results of this section is [247], including proofs of most of the results given here.

Lemma 3.4.1 *Let K be a convex subset of a real inner product space V. For any $u \in V$, $\widehat{u} \in K$ is its best approximation in K if and only if it satisfies*

$$(u - \widehat{u}, v - \widehat{u}) \leq 0 \quad \forall\, v \in K. \tag{3.4.1}$$

Proof. Suppose $\widehat{u} \in K$ is a best approximation of u. Let $v \in K$ be arbitrary. Then, since K is convex, $\widehat{u} + \lambda\,(v - \widehat{u}) \in K$, $\lambda \in [0, 1]$. Hence the function

$$\varphi(\lambda) = \|u - [\widehat{u} + \lambda\,(v - \widehat{u})]\|^2, \quad \lambda \in [0, 1],$$

has its minimum at $\lambda = 0$. We then have

$$0 \leq \varphi'(0) = -2\,(u - \widehat{u}, v - \widehat{u}),$$

i.e., (3.4.1) holds.

Conversely, assume (3.4.1) is valid. Then for any $v \in K$,

$$\begin{aligned} \|u - v\|^2 &= \|(u - \widehat{u}) + (\widehat{u} - v)\|^2 \\ &= \|u - \widehat{u}\|^2 + 2\,(u - \widehat{u}, \widehat{u} - v) + \|\widehat{u} - v\|^2 \\ &\geq \|u - \widehat{u}\|^2, \end{aligned}$$

i.e., $\widehat{u}$ is a best approximation of u in K. □

The geometric interpretation of (3.4.1) is that for any $v \in K$, the angle between the two vectors $u - \widehat{u}$ and $v - \widehat{u}$ is in the range $[\pi/2, \pi]$.

Corollary 3.4.2 *Let K be a convex subset of an inner product space V. Then for any $u \in V$, its best approximation is unique.*

Proof. Assume both $\widehat{u}_1, \widehat{u}_2 \in K$ are best approximations. Then from Lemma 3.4.1,

$$(u - \widehat{u}_1, v - \widehat{u}_1) \leq 0 \quad \forall\, v \in K.$$

In particular, we choose $v = \widehat{u}_2$ to obtain

$$(u - \widehat{u}_1, \widehat{u}_2 - \widehat{u}_1) \leq 0.$$

Similarly,

$$(u - \widehat{u}_2, \widehat{u}_1 - \widehat{u}_2) \leq 0.$$

Adding the last two inequalities, we get

$$\|\widehat{u}_1 - \widehat{u}_2\|^2 \leq 0.$$

Therefore, $\widehat{u}_1 = \widehat{u}_2$. □

Now combining the above uniqueness result and the existence results from the previous subsection, we can state the following theorem.

Theorem 3.4.3 *Assume $K \subset V$ is a non-empty, closed, convex subset of a Hilbert space V. Then for any $u \in V$, there is a unique element $\widehat{u} \in K$ such that*

$$\|u - \widehat{u}\| = \inf_{v \in K} \|u - v\|.$$

The element $\widehat{u}$ is also characterized by the inequality (3.4.1).

Actually, Theorem 3.4.3 is usually proved directly in most textbooks on functional analysis by employing the inner product structure of the space V. Let $\{u_n\}_{n \geq 1} \subset K$ be a minimizing sequence:

$$\|u - u_n\| \to \alpha \equiv \inf_{v \in K} \|u - v\|.$$

Then by the Parallelogram Law (1.3.3), we have

$$\|2\,u - u_n - u_m\|^2 + \|u_n - u_m\|^2 = 2\left(\|u - u_n\|^2 + \|u - u_m\|^2\right),$$

i.e.

$$\|u_n - u_m\|^2 = 2\left(\|u - u_n\|^2 + \|u - u_m\|^2\right) - 4\,\|u - (u_n + u_m)/2\|^2.$$

Since K is convex, $(u_n + u_m)/2 \in K$ and $\|u - (u_n + u_m)/2\| \geq \alpha$. Therefore,

$$\|u_n - u_m\|^2 \leq 2\left(\|u - u_n\|^2 + \|u - u_m\|^2\right) - 4\,\alpha^2.$$

So $\{u_n\}_{n \geq 1}$ is a Cauchy sequence. Since V is complete, the sequence has a (strong) limit $\widehat{u} \in V$, which belongs to K due to the closedness of K. It is easy to see that the limit is the best approximation.

We call $\widehat{u}$ the projection of u onto the closed convex set K, and write $\widehat{u} = P_K(u)$. The operator $P_K : V \to K$ is called the *projection operator* of V onto K. In general, P_K is a nonlinear operator. However, when K is a closed subspace, the projection operator is linear; see Theorem 3.4.7 below. It is not difficult to prove the following properties of the projection operator by using the characterization (3.4.1).

Proposition 3.4.4 *Assume $K \subset V$ is a non-empty, closed, convex subset of a Hilbert space V. Then the projection operator is monotone:*

$$(P_K(u) - P_K(v), u - v) \geq 0 \quad \forall\, u, v \in V,$$

and non-expansive:

$$\|P_K(u) - P_K(v)\| \leq \|u - v\| \quad \forall\, u, v \in V.$$

We have the following two variants of Theorem 3.4.3.

Theorem 3.4.5 *Assume $K \subset V$ is a convex and closed finite-dimensional subset of an inner product space V. Then for any $u \in V$, there is a unique element $\widehat{u} \in K$ such that*

$$\|u - \widehat{u}\| = \inf_{v \in K} \|u - v\|.$$

Theorem 3.4.6 *Assume K is a complete subspace of an inner product space V. Then for any $u \in V$, there is a unique element $\widehat{u} \in K$ such that*

$$\|u - \widehat{u}\| = \inf_{v \in K} \|u - v\|.$$

Moreover, the best approximation $\widehat{u} \in K$ is characterized by the property

$$(u - \widehat{u}, v) = 0 \quad \forall\, v \in K. \tag{3.4.2}$$

Theorem 3.4.6 does not follow from the existence theorems given in the previous section. Nevertheless, it can be shown with the idea given after Theorem 3.4.3.

A proof of the equivalence between (3.4.1) and (3.4.2) for the case of a subspace K is left as an exercise (Exercise 3.4.8).

A geometric interpretation of the characterization (3.4.2) is that the "error" $u - \widehat{u}$ is orthogonal to the subspace K. The projection mapping P_K is then called an *orthogonal projection operator*. Its main properties are summarized in the next theorem. For a detailed discussion, see [135, pp. 147, 172–174].

Theorem 3.4.7 *Assume K is a complete subspace of an inner product space V. Then the orthogonal projection operator $P_K : V \to V$ is linear, self-adjoint, i.e.,*

$$(P_K u, v) = (u, P_K v) \quad \forall\, u, v \in V. \tag{3.4.3}$$

In addition,

$$\|v\|^2 = \|P_K v\|^2 + \|v - P_K v\|^2 \quad \forall\, v \in V; \tag{3.4.4}$$

and as a consequence,

$$\|P_K\| = 1. \tag{3.4.5}$$

An important special situation arises when we know an orthonormal basis $\{\phi_n\}_{n\geq 1}$ of the space V, and $K = V_n = \operatorname{span}\{\phi_1, \ldots, \phi_n\}$. The element $P_n u \in V_n$ is the minimizer of

$$\min_{v\in V_n} \|u - v\|.$$

We find this minimizer by considering the minimization of the non-negative function

$$f(b_1, \ldots, b_n) = \left\| u - \sum_{i=1}^{n} b_i \phi_i \right\|^2, \quad b_1, \ldots, b_n \in \mathbb{R},$$

which is equivalent to minimizing $\|u - v\|$ for $v \in V_n$. It is straightforward to obtain the identity

$$f(b_1, \ldots, b_n) = \|u\|^2 - \sum_{i=1}^{n} |(u, \phi_i)|^2 + \sum_{i=1}^{n} |b_i - (u, \phi_i)|^2,$$

the verification of which is left to the reader. Clearly, the minimum of f is attained by letting $b_i = (u, \phi_i)$, $i = 1, \ldots, n$. Thus the orthogonal projection of u into V_n is given by

$$P_n u = \sum_{i=1}^{n} (u, \phi_i)\, \phi_i. \tag{3.4.6}$$

Since

$$\|u - P_n u\| = \inf_{v\in V_n} \|u - v\| \to 0 \quad \text{as } n \to \infty,$$

we have the expansion

$$u = \lim_{n\to\infty} \sum_{i=1}^{n} (u, \phi_i)\, \phi_i = \sum_{i=1}^{\infty} (u, \phi_i)\, \phi_i,$$

where the limit is understood in the sense of the norm $\|\cdot\|$. The quantity $P_n u$ is also called the *least squares approximation* of u by the elements of V_n.

Example 3.4.8 A very important application is the least squares approximation of functions by polynomials. Let $V = L^2(-1, 1)$, and $V_n = \mathbb{P}_n(-1, 1)$ the space of polynomials of degree less than or equal to n. Note that the dimension of V_n is $n + 1$ instead of n; this fact does not have any essential influence in the above discussions. An orthonormal polynomial basis for V is known, $\{\phi_n \equiv L_n\}_{n\geq 0}$ consists of the normalized Legendre polynomials,

$$L_n(x) = \sqrt{\frac{2\,n + 1}{2}}\, \frac{1}{2^n n!}\, \frac{d^n}{dx^n} \left[(x^2 - 1)^n \right], \quad n \geq 0, \tag{3.4.7}$$

where we apply the convention that $0! = 1$, $d^0 f(x)/dx^0 = f(x)$. For any $u \in V$, its least squares best approximation from $\mathbb{P}_n(-1,1)$ is given by the formula

$$P_n u(x) = \sum_{i=0}^{n} (u, L_i)_{L^2(-1,1)} \, L_i(x).$$

We have the convergence

$$\lim_{n\to\infty} \|u - P_n u\|_{L^2(-1,1)} = 0.$$

Therefore,

$$\begin{aligned} \|u\|^2_{L^2(-1,1)} &= \lim_{n\to\infty} \|P_n u\|^2_{L^2(-1,1)} \\ &= \lim_{n\to\infty} \sum_{i=0}^{n} |(u, L_i)_{L^2(-1,1)}|^2 \\ &= \sum_{i=0}^{\infty} |(u, L_i)_{L^2(-1,1)}|^2, \end{aligned}$$

known as Parseval's equality. We also have

$$u = \lim_{n\to\infty} \sum_{i=0}^{n} (u, L_i)_{L^2(-1,1)} \, L_i = \sum_{i=0}^{\infty} (u, L_i)_{L^2(-1,1)} \, L_i$$

in the sense of $L^2(-1,1)$ norm. □

Example 3.4.9 An equally important example is the least squares approximation of a function $f \in L^2(0, 2\pi)$ by trigonometric polynomials (see (3.2.13)). Let $V_n = \mathbb{T}_n$, the set of all trigonometric polynomials of degree $\le n$. Then the least squares approximation is given by

$$p_n(x) = \frac{1}{2} a_0 + \sum_{j=1}^{n} \left[a_j \cos(jx) + b_j \sin(jx) \right], \tag{3.4.8}$$

where

$$\begin{aligned} a_j &= \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(jx) \, dx, \quad j \ge 0, \\ b_j &= \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(jx) \, dx, \quad j \ge 1. \end{aligned} \tag{3.4.9}$$

As with Example 3.4.8, we can look at the convergence of (3.4.8). This leads to the well-known Fourier series expansion

$$f(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} \left[a_j \cos(jx) + b_j \sin(jx) \right]. \tag{3.4.10}$$

Further development of this example is left as Exercise 3.4.10. □

Exercise 3.4.1 Let $V = L^2(\Omega)$, and for $r > 0$,

$$K = \{v \in V \mid \|v\|_{L^2(\Omega)} \leq r\}.$$

For any $u \in V$, find its projection on K.

Exercise 3.4.2 Consider a subset in $V = L^2(\Omega) \times L^2(\Omega)$:

$$K = \{\boldsymbol{v} = (v_1, v_2) \in V \mid v_1 \geq v_2 \text{ a.e. in } \Omega\}.$$

Show that K is non-empty, closed and convex. For any $\boldsymbol{u} = (u_1, u_2) \in V$, verify that its projection onto K is $\boldsymbol{w} = (w_1, w_2) \in K$ with

$$w_1 = u_1 + (u_2 - u_1)_+/2, \quad w_2 = u_2 - (u_2 - u_1)_+/2.$$

Here, $t_+ = \max\{t, 0\}$.

Exercise 3.4.3 Show directly that if K is a non-empty closed convex set in a Hilbert space V, then there is a unique element $u \in K$ satisfying

$$\|u\| = \inf_{v \in K} \|v\|.$$

Exercise 3.4.4 Let V be an inner product space. A non-zero vector $u_0 \in V$ generates a one-dimensional subspace

$$U_0 = \{\alpha\, u_0 \mid \alpha \in \mathbb{R}\}.$$

For any $u \in V$, find its distance to U_0:

$$\operatorname{dist}(u, U_0) = \inf\{\|u - v\| \mid v \in U_0\}.$$

Exercise 3.4.5 Corollary 3.4.2 can be proved directly. Suppose both $\hat{u}_1, \hat{u}_2 \in K$ are best approximations to $u \in V$. Then due to the convexity of K, $(\hat{u}_1 + \hat{u}_2)/2 \in K$ and is also a best approximation to u:

$$\begin{aligned} \|u - (\hat{u}_1 + \hat{u}_2)/2\| &= \|(u - \hat{u}_1)/2 + (u - \hat{u}_2)/2\| \\ &\leq \|u - \hat{u}_1\|/2 + \|u - \hat{u}_2\|/2 \\ &= \|u - \hat{u}_1\|. \end{aligned}$$

The inequality must be an equality, and so $(u - \hat{u}_1)$ or $(u - \hat{u}_2)$ is a nonnegative multiple of the other. Then it can be shown that $\hat{u}_1 = \hat{u}_2$. Carry out the detail of this argument.

Exercise 3.4.6 Another proof of Corollary 3.4.2 is through an application of the Parallelogram Law (1.3.3). Give the proof by writing

$$\begin{aligned} \|\hat{u}_1 - \hat{u}_2\|^2 &= \|(u - \hat{u}_1) - (u - \hat{u}_2)\|^2 \\ &= 2\,\|u - \hat{u}_1\|^2 + 2\,\|u - \hat{u}_2\|^2 - 4\,\|u - (\hat{u}_1 + \hat{u}_2)/2\|^2, \end{aligned}$$

and noting that $(\hat{u}_1 + \hat{u}_2)/2 \in K$.

Exercise 3.4.7 Prove Proposition 3.4.4.

Exercise 3.4.8 Show that when K is a subspace, the characterization (3.4.1) is equivalent to (3.4.2).

Exercise 3.4.9 Prove Theorem 3.4.7.

Exercise 3.4.10 Given a function $f \in L^2(0, 2\pi)$, show that its best approximation in the space $\mathbb{T}_n$ with respect to the norm of $L^2(0, 2\pi)$ is given by the partial sum

$$\frac{a_0}{2} + \sum_{j=1}^{n} (a_j \cos jx + b_j \sin jx)$$

of the Fourier series of f with the Fourier coefficients given by (3.4.9). Derive Parseval's equality for this case.

Exercise 3.4.11 Repeat Exercise 3.4.10, but use the basis

$$\left\{ e^{ijx} \;\middle|\; -n \le j \le n \right\}$$

for T_n. Find a formula for the least squares approximation of $f(x)$ in $L^2(0, 2\pi)$. Give Parseval's equality and give a formula for $\|u - P_n u\|$ in terms of the Fourier coefficients of f when using this basis.

Exercise 3.4.12 Toeplitz matrices and their specialization, circulant matrices, appear in imaging problems and other applications ([227]). An $n \times n$ Toeplitz matrix is of the form

$$T = \begin{pmatrix} t_0 & t_{-1} & \cdots & t_{2-n} & t_{1-n} \\ t_1 & t_0 & \cdots & t_{3-n} & t_{2-n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ t_{n-2} & t_{n-3} & \cdots & t_0 & t_{-1} \\ t_{n-1} & t_{n-2} & \cdots & t_1 & t_0 \end{pmatrix},$$

which is denoted as $T = \text{toeplitz}(t_{1-n}, \ldots, t_0, \ldots, t_{n-1})$. Here $t_j \in \mathbb{C}$, $1 - n \le j \le n - 1$. An $n \times n$ circulant matrix is a special kind of Toeplitz matrix: $C = \text{toeplitz}(c_1, \ldots, c_{n-1}, c_0, c_1, \ldots, c_{n-1})$, written as $C = \text{circulant}(c_0, \ldots, c_{n-1})$. Let $\mathbb{C}_c^{n\times n} \subset \mathbb{C}^{n\times n}$ be the subspace of all the $n \times n$ circulant matrices. In $\mathbb{C}^{n\times n}$, we use the Frobenius inner product

$$(A, B)_F = \text{tr}(B^H A), \quad A, B \in \mathbb{C}^{n\times n}.$$

Define the circulant right shift matrix $R = \text{circulant}(0, 1, 0, \ldots, 0)$.
(a) Derive the formula

$$C = \sum_{j=0}^{n-1} c_j R^j.$$

(b) Show that $R^j/\sqrt{n}$, $0 \le j \le n - 1$, form an orthonormal basis of the subspace $\mathbb{C}_c^{n\times n}$.
(c) Given $A \in \mathbb{R}^{n\times n}$, the best circulant approximation to A,

$$C(A) = \arg\min_{C \in \mathbb{C}_c^{n\times n}} \|C - A\|_F,$$

is $C(A) = \text{circulant}(c_0, \ldots, c_{n-1})$, with $c_j = (A, R^j)_F/n$, $0 \le j \le n-1$. In particular, for a real Toeplitz matrix $T = \text{toeplitz}(t_{1-n}, \ldots, t_0, \ldots, t_{n-1})$, $C(T) = \text{circulant}(c_0, \ldots, c_{n-1})$, with $c_j = [(n-j)t_j + jt_{j-n}]/n$, $0 \le j \le n-1$.

Circulant approximations to square matrices can be useful in constructing preconditioners for Toeplitz systems.

3.5 Orthogonal polynomials

The discussion of Example 3.4.8 at the end of the previous section can be extended in a more general framework of weighted L^2-spaces. As in Example 3.4.8, we use the interval $[-1, 1]$; all other finite intervals $[a, b]$ can be converted to $[-1, 1]$ by a simple linear change of variables. Let $w(x)$ be a weight function on $[-1, 1]$, i.e. it is positive almost everywhere and it is integrable on $[-1, 1]$. Then we can introduce a weighted function space

$$L^2_w(-1,1) = \left\{ v \text{ is measurable on } [-1,1] \;\middle|\; \int_{-1}^{1} |v(x)|^2 w(x)\, dx < \infty \right\}. \tag{3.5.1}$$

This is a Hilbert space with the inner product

$$(u, v)_{0,w} = \int_{-1}^{1} u(x)\, v(x)\, w(x)\, dx$$

and the corresponding norm

$$\|v\|_{0,w} = \sqrt{(v, v)_{0,w}}\,.$$

Two functions $u, v \in L^2_w(-1, 1)$ are said to be orthogonal if $(u, v)_{0,w} = 0$.

Starting with the monomials $\{1, x, x^2, \ldots\}$, we can apply the Gram-Schmidt procedure described in Section 1.3 to construct a system of orthogonal polynomials $\{p_n(x)\}_{n=0}^{\infty}$ such that the degree of p_n is n. For any $u \in L^2_w(-1, 1)$, the best approximating polynomial of degree less than or equal to N is

$$P_N u(x) = \sum_{n=0}^{N} \xi_n p_n(x), \quad \xi_n = \frac{(u, p_n)_{0,w}}{\|p_n\|^2_{0,w}}, \quad 0 \le n \le N. \tag{3.5.2}$$

This can be verified directly. The best approximation $P_N u$ is characterized by the property that it is the orthogonal projection of u onto the polynomial space $\mathbb{P}_N(-1, 1)$ with respect to the inner product $(\cdot, \cdot)_{0,w}$.

A family of well-known orthogonal polynomials, called the *Jacobi polynomials*, are related to the weight function

$$w^{(\alpha,\beta)}(x) = (1-x)^\alpha (1+x)^\beta, \quad -1 < \alpha, \beta < 1. \tag{3.5.3}$$

A detailed discussion of these polynomials can be found in the reference [220]. Here we mention some results for two of the most important special cases.

When $\alpha = \beta = 0$, the Jacobi polynomials become Legendre polynomials, which were discussed in Example 3.4.8. Conventionally, the Legendre polynomials are defined to be

$$L_0(x) = 1, \quad L_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left[(x^2 - 1)^n \right], \ n \geq 1. \tag{3.5.4}$$

These polynomials are orthogonal, and

$$(L_m, L_n)_0 = \frac{2}{2\,n + 1}\, \delta_{mn}. \tag{3.5.5}$$

The Legendre polynomials satisfy the differential equation

$$\left[(1 - x^2)\, L_n'(x) \right]' + n\,(n+1)\, L_n(x) = 0, \quad n = 0, 1, \ldots,$$

and the triple recursion formula

$$L_{n+1}(x) = \frac{2\,n+1}{n+1}\, x\, L_n(x) - \frac{n}{n+1}\, L_{n-1}(x), \quad n = 1, 2, \ldots \tag{3.5.6}$$

with $L_0(x) = 1$ and $L_1(x) = x$. Graphs of normalized Legendre polynomials of degrees $n = 0, 1, 2, 3$ were given earlier in Figure 1.2 of Subsection 1.3.2 in Chapter 1.

To present some error estimates related to orthogonal projection polynomials, we need to use the notion of Sobolev spaces as is reviewed in Chapter 7. A reader without prior knowledge on Sobolev spaces may skip the following error estimates in a first time reading.

For any $u \in L^2(-1, 1)$, its N-th degree $L^2(-1, 1)$-projection polynomial $P_N u$ is

$$P_N u(x) = \sum_{n=0}^{N} \xi_n L_n(x), \quad \xi_n = \frac{2\,n+1}{2}\, (u, L_n)_0, \quad 0 \leq n \leq N.$$

It is shown in [46] that if $u \in H^s(-1, 1)$ with $s > 0$, then the following error estimates hold:

$$\begin{aligned} \|u - P_N u\|_0 &\leq c\, N^{-s}\, \|u\|_s, \\ \|u - P_N u\|_1 &\leq c\, N^{3/2-s}\, \|u\|_s. \end{aligned}$$

Here $\|\cdot\|_s$ denotes the $H^s(-1, 1)$-norm, and below we use $(\cdot, \cdot)_1$ for the inner product in $H^1(-1, 1)$.

Notice that the error estimate in the $L^2(-1, 1)$-norm is of optimal order as expected, yet the error estimate in the $H^1(-1, 1)$-norm is not of optimal

order. In order to improve the approximation order also in the $H^1(-1,1)$-norm, another orthogonal projection operator $P_{1,N} : H^1(-1,1) \to \mathbb{P}_N$ can be introduced: For $u \in H^1(-1,1)$, its projection $P_{1,N}u \in \mathbb{P}_N$ is defined by

$$(P_{1,N}u, v)_1 = (u, v)_1 \quad \forall\, v \in \mathbb{P}_N.$$

It is shown in [162] that

$$\|u - P_{1,N}u\|_k \le c\, N^{k-s} \|u\|_s, \quad k = 0, 1,\ s \ge 1. \tag{3.5.7}$$

Notice that the error is of optimal order in both the $L^2(-1,1)$-norm and the $H^1(-1,1)$-norm.

Another important special case of (3.5.3) is when $\alpha = \beta = -1/2$. The weight function here is

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

and the weighted inner product is

$$(u, v)_{0,w} = \int_{-1}^{1} \frac{u(x)\, v(x)}{\sqrt{1-x^2}}\, dx.$$

The corresponding orthogonal polynomials are called *Chebyshev polynomials of the first kind*,

$$T_n(x) = \cos(n \arccos x), \quad n = 0, 1, \ldots. \tag{3.5.8}$$

These functions are orthogonal,

$$(T_m, T_n)_{0,w} = \frac{\pi}{2}\, c_n \delta_{mn}, \quad n, m \ge 0$$

with $c_0 = 2$ and $c_n = 1$ for $n \ge 1$. The Chebyshev polynomials satisfy the differential equation

$$-\left[\sqrt{1-x^2}\, T_n'(x)\right]' = n^2 \frac{T_n(x)}{\sqrt{1-x^2}}, \quad n = 0, 1, \ldots$$

and the triple recursion formula

$$T_{n+1}(x) = 2\, x\, T_n(x) - T_{n-1}(x), \quad n \ge 1, \tag{3.5.9}$$

with $T_0(x) = 1$ and $T_1(x) = x$.

Above, we considered orthogonal polynomials defined on the interval $[-1,1]$. On a general finite interval $[a,b]$ $(b > a)$, we can use a simple linear transformation of the independent variables, and reduce the study of orthogonal polynomials on the interval $[a,b]$ to that on the interval $[-1,1]$. It is also possible to study orthogonal polynomials defined on unbounded intervals.

Orthogonal polynomials are important in the derivation and analysis of Gaussian numerical integration (see e.g., [15, Section 5.3]), and in the study of a family of powerful numerical methods, called spectral methods, for solving differential equations (see e.g., [96, 45, 35]). For a more extended introduction to orthogonal polynomials, see [4, Chaps. 5–7], [64, Chap. 10], [88]. Additional results for orthogonal polynomials and polynomial approximation are given later in Section 3.7.2. In Chapter 14, an introduction is given to multivariable orthogonal polynomials and multivariable polynomial approximation theory.

Exercise 3.5.1 Use (3.5.4) and integration by parts to show $(L_m, L_n) = 0$ for $m \neq n$, $m, n \geq 0$.
Hint: Assume $n > m$. Use repeated integration by parts in evaluating (L_m, L_n). Note that $\left(x^2 - 1\right)^n$ has the roots 1 and -1, both of multiplicity n.

Exercise 3.5.2 Use (3.5.4) and integration by parts to show

$$(L_n, L_n)_0 = \frac{2}{2\,n+1}.$$

Exercise 3.5.3 Derive formulas for the Legendre polynomials of (3.5.4) and the Chebyshev polynomials of (3.5.8) over a general interval $[a, b]$. For the Chebyshev polynomials, what is the appropriate weight function over $[a, b]$?

Exercise 3.5.4 Derive (3.5.9) from (3.5.8).

Exercise 3.5.5 The existence of a three-term recursion formula for orthogonal polynomials is not a coincidence for the Legendre polynomials (see (3.5.6)) and the Chebyshev polynomials (see (3.5.9)). In general, let $p_0, p_1, p_2, \cdots$ be a sequence of orthogonal polynomials with respect to the inner product

$$(u, v)_{0,w} = \int_a^b u(x)\, v(x)\, w(x)\, dx$$

with a weight function w, and $\deg p_n = n$, $n \geq 0$. Prove the following three-term recursion formula:

$$p_{n+1}(x) = (a_n x + b_n)\, p_n(x) + c_n p_{n-1}(x), \quad n \geq 1$$

for suitable constants a_n, b_n and c_n.
Hint: Write

$$x\, p_n(x) = \sum_{i=0}^{n+1} \alpha_i p_i(x).$$

Then,

$$\alpha_j \|p_j\|_{0,w}^2 = (x\, p_n(x), p_j(x))_{0,w}, \quad j = 0, 1, \cdots, n+1.$$

Show that for $j \leq n-2$,

$$(x\, p_n(x), p_j(x))_{0,w} = 0.$$

Exercise 3.5.6 As a continuation of Exercise 3.5.5, find the following formulas for a_n, b_n and c_n. Assume the orthogonal polynomials $\{p_n(x)\}$ have the form

$$p_n(x) = A_n x^n + B_n x^{n-1} + \cdots .$$

Then

$$a_n = \frac{A_{n+1}}{A_n}, \quad b_n = a_n \left(\frac{B_{n+1}}{A_{n+1}} - \frac{B_n}{A_n} \right), \quad c_n = -\frac{A_{n+1} A_{n-1}}{A_n^2} \cdot \frac{\gamma_n}{\gamma_{n-1}}$$

with $\gamma_n = (p_n, p_n)$, $n \geq 0$.

Exercise 3.5.7 Use (3.5.4), (3.5.5) and Exercise 3.5.6 to obtain the triple recursion relation in (3.5.6) for the Legendre polynomials $L_n(x)$, $n \geq 0$.

Exercise 3.5.8 Find the zeros of $T_n(x)$ for $n \geq 1$. Find the points at which

$$\max_{-1 \leq x \leq 1} |T_n(x)|$$

is attained.

Exercise 3.5.9 Using the Gram-Schmidt process, construct orthogonal polynomials of degrees 0, 1, 2 for the weight function $w(x) = \log(1/x)$ on $[0, 1]$.

Exercise 3.5.10 For $n \geq 0$, define

$$S_n(x) = \frac{1}{n+1} T'_{n+1}(x)$$

using the Chebyshev polynomials of (3.5.8). These new polynomials $\{S_n(x)\}$ are called *Chebyshev polynomials of the second kind*.
(a) Show that $\{S_n(x)\}$ is an orthogonal family on $[-1, 1]$ with respect to the weight function $w(x) = \sqrt{1 - x^2}$.
(b) Show that $\{S_n(x)\}$ also satisfies the triple recursion relation (3.5.9).

Exercise 3.5.11 Lobatto polynomials are defined as follows:

$$\ell_0(x) = \frac{1 - x}{2}, \quad \ell_1(x) = \frac{1 + x}{2},$$
$$\ell_k(x) = \frac{1}{\|L_{k-1}\|_{L^2(-1,1)}} \int_{-1}^{x} L_{k-1}(t)\, dt, \quad k \geq 2,$$

where L_{k-1}, $k \geq 2$, are the Legendre polynomials. Recall that $\|L_{k-1}\|_{L^2(-1,1)} = \sqrt{2/(2k-1)}$.
(a) Find formulas for $\ell_2(x)$, $\ell_3(x)$ and $\ell_4(x)$.
(b) Show that for any integer $p > 0$, $\{\ell_k\}_{0 \leq k \leq p}$ forms a basis for the polynomial space $\mathbb{P}_p$.
(c) For $k \geq 2$, $\ell_k(-1) = \ell_k(1) = 0$.
(d) For $i \neq j$ with $\max\{i, j\} > 1$,

$$\int_{-1}^{1} \ell'_i(x)\, \ell'_j(x)\, dx = 0.$$

Lobatto polynomials are used in the design of hierarchic shape functions in the p-version finite element method (see e.g., [209]). The orthogonality (d) is the property that makes these polynomials especially useful in this context.

3.6 Projection operators

In Section 3.4, we have introduced the notion of a projection operator in an inner product space. In this section, we consider projection operators on subspaces of more general linear spaces or normed spaces. Projection operators are useful in discussing many approximation methods. Intuitively we are approximating elements of a vector space V using elements of a subspace W. Originally, this generalized the construction of an orthogonal projection from Euclidean geometry, finding the orthogonal projection of an element $v \in V$ in the subspace W. This has since been extended to general linear spaces which do not possess an inner product; and hence our discussion approaches the definition of projection operators from another perspective.

Definition 3.6.1 *Let V be a linear space, and V_1 and V_2 be subspaces of V. We say V is the* direct sum *of V_1 and V_2 and write $V = V_1 \oplus V_2$, if any element $v \in V$ can be uniquely decomposed as*

$$v = v_1 + v_2, \quad v_1 \in V_1,\ v_2 \in V_2. \tag{3.6.1}$$

Furthermore, if V is an inner product space, and $(v_1, v_2) = 0$ for any $v_1 \in V_1$ and any $v_2 \in V_2$, then V is called the orthogonal direct sum *of V_1 and V_2.*

There exists a one-to-one correspondence between direct sums and linear operators P satisfying $P^2 = P$.

Proposition 3.6.2 *Let V be a linear space. Then $V = V_1 \oplus V_2$ if and only if there is a linear operator $P : V \to V$ with $P^2 = P$ such that in the decomposition* (3.6.1)*, $v_1 = Pv$, $v_2 = (I - P)v$, and moreover, $V_1 = P(V)$ and $V_2 = (I - P)(V)$.*

Proof. Let $V = V_1 \oplus V_2$. Then $Pv = v_1$ defines an operator from V to V. It is easy to verify that P is linear and maps V onto V_1 ($Pv_1 = v_1\ \forall\, v_1 \in V_1$), and so $V_1 = P(V)$. Obviously $v_2 = (I - P)v$ and $(I - P)v_2 = v_2\ \forall\, v_2 \in V_2$.

Conversely, with the operator P, for any $v \in V$ we have the decomposition $v = Pv + (I - P)v$. We must show this decomposition is unique. Suppose $v = v_1 + v_2$, $v_1 \in V_1$, $v_2 \in V_2$. Then $v_1 = Pw$ for some $w \in V$. This implies $Pv_1 = P^2w = Pw = v_1$. Similarly, $Pv_2 = 0$. Hence, $Pv = v_1$, and then $v_2 = v - v_1 = (I - P)v$. □

Definition 3.6.3 *Let V be a Banach space. An operator $P \in \mathcal{L}(V)$ with the property $P^2 = P$ is called a* projection operator. *The subspace $P(V)$ is called the corresponding projection space. The direct sum*

$$V = P(V) \oplus (I - P)(V)$$

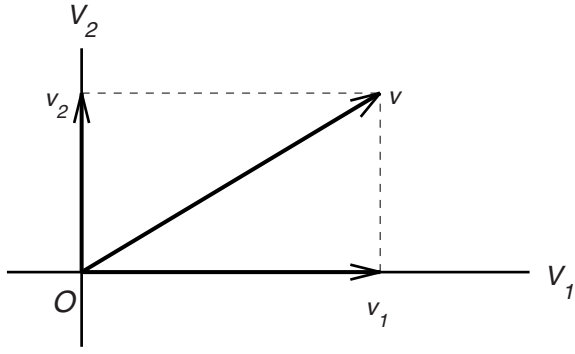

FIGURE 3.4. Orthogonal projection in $\mathbb{R}^2$

is called a topological direct sum.

If V is a Hilbert space, P is a projection operator, and $V = P(V) \oplus (I - P)(V)$ is an orthogonal direct sum, then we call P an orthogonal projection operator.

It is easy to see that a projection operator P is orthogonal if and only if

$$(Pv, (I - P)w) = 0 \quad \forall\, v, w \in V. \tag{3.6.2}$$

Example 3.6.4 Figure 3.4 illustrates the orthogonal direct decomposition of an arbitrary vector in $\mathbb{R}^2$ which defines an orthogonal projection operator P from $\mathbb{R}^2$ to V_1. In particular, when V_1 is the x_1-axis, we have

$$Pv = \begin{pmatrix} v_1 \\ 0 \end{pmatrix} \quad \text{for } v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \qquad \square$$

Example 3.6.5 (Lagrange interpolation) Let $V = C[a, b]$, $V_1 = \mathbb{P}_n$ the space of the polynomials of degree less than or equal to n, and let $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ be a partition of the interval $[a, b]$. For $v \in C[a, b]$, we define $Pv \in \mathbb{P}_n$ to be the Lagrange interpolant of v corresponding to the partition Δ, i.e., Pv satisfies the interpolation conditions: $Pv(x_i) = v(x_i)$, $0 \le i \le n$. From the discussion of Section 3.2, the interpolant Pv is uniquely determined. The uniqueness of the interpolant implies that P is a projection operator. Explicitly,

$$Pv(x) = \sum_{i=0}^{n} \left(\prod_{j \neq i} \frac{x - x_j}{x_i - x_j} \right) v(x_i),$$

using the Lagrange formula for the interpolant. $\square$

Example 3.6.6 (Piecewise linear interpolation) Again we let $V = C[a,b]$ and $\Delta : a = x_0 < x_1 < \cdots < x_n = b$ a partition of the interval $[a,b]$. This time, we take V_1 to be the space of continuous piecewise linear functions:

$$V_1 = \{v \in C[a,b] \mid v|_{[x_{i-1},x_i]} \text{ is linear, } 1 \le i \le n\}.$$

Then for any $v \in C[a,b]$, Pv is the piecewise linear function uniquely determined by the interpolation conditions $Pv(x_i) = v(x_i)$, $0 \le i \le n$. This is an example of a finite element space. □

Example 3.6.7 Let V_n be an n-dimensional subspace of a Hilbert space V. Suppose $\{u_1, \ldots, u_n\}$ is an orthonormal basis of V_n. For any $v \in V$, the formula

$$Pv = \sum_{i=1}^{n} (u_i, v)\, u_i$$

defines an orthogonal projection from V onto V_n. □

Example 3.6.8 Recall the least squares approximation using trigonometric polynomials, in (3.4.8)–(3.4.9). This defines an orthogonal projection from $L^2(0, 2\pi)$ to $\mathbb{T}_n$. Denote it by $\mathcal{F}_n f$. In the following section, we discuss $\mathcal{F}_n$ as a projection from $C_p(2\pi)$ to $\mathbb{T}_n$. Recall from Examples 1.1.2 (g), 1.2.5 (a) that the space $C_p(2\pi)$ consists of all continuous functions g on $\mathbb{R}$ for which

$$g(x + 2\pi) \equiv g(x)$$

and the norm is $\|\cdot\|_\infty$. Proposition 3.6.9, given below, also applies to $\mathcal{F}_n$ and the linear space $L^2(0, 2\pi)$. □

If V is an inner product space, then we define the orthogonal complement of a subspace V_1 as

$$V_1^\perp = \{v \in V \mid (v, v_1) = 0 \quad \forall\, v_1 \in V_1\}.$$

The proof of the following is left as Exercise 3.6.6.

Proposition 3.6.9 (Orthogonal projection) *Let V_1 be a closed linear subspace of the Hilbert space V, with its orthogonal complement $V_1^\perp$. Let $P : V \to V_1$. Then:*
(a) *The operator P is an orthogonal projection if and only if it is a self-adjoint projection.*
(b) *Each orthogonal projection P is continuous, $\|P\| \le 1$, and $\|P\| = 1$ for $P \ne 0$.*
(c) $V = V_1 \oplus V_1^\perp$.
(d) *There exists exactly one orthogonal projection operator P from V onto V_1. We have*

$$\|v - Pv\| = \inf_{w \in V_1} \|v - w\| \quad \forall\, v \in V.$$

The operator $I - P$ is the orthogonal projection onto $V_1^{\perp}$.
(e) *If $P : V \to V$ is an orthogonal projection operator, then $P(V)$ is a closed subspace of V, and we have the orthogonal direct sum*

$$V = P(V) \oplus (I - P)(V).$$

Projection operators are used in defining projection methods in solving operator equations; see Section 12.1 on projection methods for integral equations.

Exercise 3.6.1 Show that if P is a projection operator (or an orthogonal projection operator), then so is $I - P$. Moreover, the range of P is the null space of $(I - P)$, and the range of $(I - P)$ is the null space of P.

Exercise 3.6.2 Let $V = L^2(-1, 1)$, and $V_1 = \{v \in V \mid v(x) = 0 \text{ a.e. in } (-1, 0)\}$. Determine the orthogonal complement of V_1 in V.

Exercise 3.6.3 Compute the orthogonal projection in $L^2(0, 1)$ of $u(x) = e^x$ onto the subspace of all the linear functions.

Exercise 3.6.4 Show that the range of a projection operator coincides with the set of its fixed-points. A vector v is said to be a fixed-point of P if $Pv = v$.

Exercise 3.6.5 Let V be a Hilbert space, let V_1 be a finite dimensional subspace with basis $\{\varphi_1, \ldots, \varphi_n\}$, and let P be an orthogonal projection of V onto V_1. Show that $Pv = 0$ for any $v \in V$ if and only if $(v, \varphi_j) = 0$ for $j = 1, \ldots, n$.

Exercise 3.6.6 Prove Proposition 3.6.9.

Exercise 3.6.7 Let $P \neq 0$ be a bounded projection on the Banach space V. Show that $\|P\| \geq 1$. If V is a Hilbert space, and if P is an orthogonal projection, show that $\|P\| = 1$.

Exercise 3.6.8 (a) Find a formula for $\|P\|$ in Example 3.6.5.
(b) Find a formula for $\|P\|$ in Example 3.6.6.

Exercise 3.6.9 Extend Example 3.6.6 to piecewise quadratic interpolation. Use evenly spaced node points. What is $\|P\|$ in this case?

3.7 Uniform error bounds

Approximation in the uniform norm is quite important in numerical analysis and applied mathematics, and polynomials are the most important type of approximants. In this section, we focus on uniform approximation

of continuous functions of one variable. Recall from the Weierstrass theorem (Theorem 3.1.1) that any continuous function f on a bounded closed interval can be approximated uniformly by polynomials.

For many uses of approximation theory in numerical analysis, we need error bounds for the best uniform approximation of a function $f(x)$ on an interval $[a, b]$. We are interested in two such problems: the uniform approximation of a smooth function by polynomials and the uniform approximation of a smooth 2π-periodic function by trigonometric polynomials. These problems were discussed previously in Subsection 3.3.4, with Theorems 3.3.19 and 3.3.20 giving the uniqueness of the best approximants for these two forms of approximation. Initially, we study the polynomial approximation problem on the special interval $[-1, 1]$, and then the results obtained extend easily to an arbitrary interval $[a, b]$ by a simple linear change of variables. We consider the approximation of a 2π-periodic function by trigonometric polynomials as taking place on the interval $[-\pi, \pi]$ in most cases, as it aids in dealing with the special cases of even and odd 2π-periodic functions.

An important first step is to note that these two problems are closely connected. Given a function $f \in C^m[-1, 1]$, introduce the function

$$g(\theta) = f(\cos\theta). \tag{3.7.1}$$

The function g is an even 2π-periodic function and it is also m-times continuously differentiable. If we examine the best approximating trigonometric polynomial $q_n(\theta)$ for any even 2π-periodic function $g(\theta)$, the uniqueness result (Theorem 3.3.20) can be used to show that $q_n(\theta)$ has the form

$$q_n(\theta) = a_0 + \sum_{j=1}^{n} a_j \cos(j\theta). \tag{3.7.2}$$

The proof is based on using the property that g is even and that the best approximation is unique; see Exercise 3.7.2.

For the best uniform trigonometric approximation of (3.7.1), given in (3.7.2), use the substitution $x = \cos\theta$ and trigonometric identities to show the existence of $p_n \in \mathbb{P}_n$ with

$$q_n(\theta) = p_n(\cos\theta)$$

and with $p_n(x)$ having the same degree as $q_n(\theta)$. Conversely, if $p \in \mathbb{P}_n$ is given, then $q_n(\theta) \equiv p_n(\cos\theta)$ can be shown to have the form (3.7.2).

Using these results,

$$\max_{0\le\theta\le\pi} |g(\theta) - q_n(\theta)| = \max_{-1\le x\le 1} |f(x) - p_n(x)| .$$

In addition, it is straightforward to show that $p_n(x)$ must be the best uniform approximation to $f(x)$ on $[-1, 1]$. If it were not, we could produce a

better uniform approximation, call it $r_n(x)$; and then $r_n(\cos\theta)$ would be a better uniform approximation to $g(\theta)$ on $[0,\pi]$, a contradiction. This equivalence allows us to concentrate on only one of our two approximating problems, that of approximating a 2π-periodic function $g(\theta)$ by a trigonometric polynomial $q_n(\theta)$. The results then transfer immediately to the ordinary polynomial approximation problem for a function $f \in C^m[-1,1]$.

As a separate result, it can also be shown that when given any 2π-periodic function $g(\theta)$, there is a corresponding function $f(x)$ of equal smoothness for which there is an equivalence between their best uniform approximations in the respective approximating spaces $\mathbb{T}_n$ and $\mathbb{P}_n$. For this construction, see [169, page 46].

We state without proof the main results. For proofs, see Meinardus [169, Section 5.5]. Recall that the notation $C_p(2\pi)$ denotes the Banach space of 2π-periodic continuous functions, with the uniform norm as the norm.

Theorem 3.7.1 (Jackson's Theorem) *Suppose the 2π-periodic function $g(\theta)$ possesses continuous derivatives up to order k. Further assume that the k^{th} derivative satisfies a Hölder condition:*

$$\left|g^{(k)}(\theta_1) - g^{(k)}(\theta_2)\right| \le M_k \left|\theta_1 - \theta_2\right|^{\alpha}, \quad -\infty < \theta_1, \theta_2 < \infty$$

for some $M_k > 0$ and some $\alpha \in (0,1]$. (We say that $g \in C_p^{k,\alpha}(2\pi)$.) Then the error in the best approximation $q_n(\theta)$ to $g(\theta)$ satisfies

$$\max_{-\infty<\theta<\infty} |g(\theta) - q_n(\theta)| \le c^{k+1} \frac{M_k}{n^{k+\alpha}} \tag{3.7.3}$$

with $c = 1 + \pi^2/2$.

Theorem 3.7.2 (Jackson's Theorem) *Suppose $f \in C^k[-1,1]$ and that the k^{th} derivative satisfies a Hölder condition:*

$$\left|f^{(k)}(x_1) - f^{(k)}(x_2)\right| \le M_k \left|x_1 - x_2\right|^{\alpha}, \quad -1 \le x_1, x_2 \le 1$$

for some $M_k > 0$ and some $\alpha \in (0,1]$. Then the error in the best approximation $p_n(x)$ to $f(x)$ satisfies

$$\max_{-1\le x\le 1} |f(x) - p_n(x)| \le d_k c^{k+1} \frac{M_k}{n^{k+\alpha}} \tag{3.7.4}$$

with $c = 1 + \pi^2/2$ and d_k any number satisfying

$$d_k \ge \frac{n^{k+\alpha}}{n(n-1)\cdots(n-k+1)(n-k)^{\alpha}}, \quad n > k.$$

Note that the right hand fraction tends to 1 as $n \to \infty$, and therefore a finite bound d_k does exist for each $k \ge 0$.

3.7.1 Uniform error bounds for L^2-approximations

The Fourier series of a function $f \in L^2(-\pi, \pi)$ is a widely-used tool in applied and computational mathematics, and as such, error bounds are needed for the convergence of the series. We return to this topic in later chapters, deriving additional error bounds for the error in the context of Sobolev spaces (see Section 7.5). But here we look at bounds based on Theorem 3.7.1. More discussion of the Fourier series is given in Section 4.1.

The Fourier series for a function $f \in L^2(-\pi, \pi)$ was given in Example 3.4.9, with the formulas (3.4.10) and (3.4.9). As introduced in Example 3.6.8, we also use the notation $\mathcal{F}_n f$ to denote the partial Fourier series of terms of degree $\leq n$. It is straightforward to obtain bounds for $f - \mathcal{F}_n f$ in $L^2(-\pi, \pi)$, once uniform error bounds are known. Simply use

$$\|g\|_2 \leq \sqrt{2\pi}\, \|g\|_\infty, \quad g \in C[-\pi, \pi],$$

and therefore

$$\|f - \mathcal{F}_n f\|_2 \leq \sqrt{2\pi}\, \|f - \mathcal{F}_n f\|_\infty, \quad f \in C[-\pi, \pi]. \tag{3.7.5}$$

The error bounds follow immediately. An alternative set of L^2-bounds is introduced in Section 7.5.

Obtaining results in the uniform norm is more difficult. Begin by using standard trigonometric identities to rewrite the formulas (3.4.8)–(3.4.9) for $\mathcal{F}_n f$ as

$$\mathcal{F}_n f(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} D_n(x - y)\, f(y)\, dy \tag{3.7.6}$$

where

$$D_n(\theta) = \frac{1}{2} + \sum_{j=1}^{n} \cos(j\theta). \tag{3.7.7}$$

For $t \notin \{2j\pi \mid j = 0, \pm 1, \pm 2, \dots\}$, we have

$$D_n(\theta) = \frac{\sin(n + 1/2)\theta}{2 \sin(\theta/2)}. \tag{3.7.8}$$

The function D_n is called the *Dirichlet kernel function*. Many results on the behaviour of the partial Fourier sums $\mathcal{F}_n f(x)$ are obtained by an examination of the formula (3.7.6).

For $f \in C_p(2\pi)$, use this formula to obtain

$$\begin{aligned} \max_x |\mathcal{F}_n f(x)| &\leq \frac{1}{\pi} \max_x \int_{-\pi}^{\pi} |D_n(x - y)|\, dy\, \|f\|_\infty \\ &= \frac{2}{\pi} \int_0^{\pi} |D_n(y)|\, dy\, \|f\|_\infty. \end{aligned}$$

The last step uses the facts that $D_n(\theta)$ is even and 2π-periodic. From this, we see

$$\mathcal{F}_n : C_p(2\pi) \to \mathbb{T}_n \subset C_p(2\pi)$$

is a bounded projection operator with

$$\|\mathcal{F}_n\| \leq L_n \equiv \frac{2}{\pi} \int_0^\pi |D_n(y)|\, dy. \tag{3.7.9}$$

By regarding (3.7.6) as defining an integral operator from $C_p(2\pi)$ to itself, it can be seen that $\|\mathcal{F}_n\| = L_n$ (see (2.2.8)).

The numbers $\{L_n\}$ are called *Lebesgue constants*, and a great deal is known about them. In particular, it is shown in Zygmund [251, Chap. 2, p. 67] that (see Exercise 3.7.4)

$$\|\mathcal{F}_n\| = L_n = \frac{4}{\pi^2} \log n + \mathcal{O}(1), \quad n \geq 1. \tag{3.7.10}$$

Thus $\{\|\mathcal{F}_n\|\}$ is an unbounded sequence. This implies the existence of a function $f \in C_p(2\pi)$ for which $\mathcal{F}_n f$ does not converge uniformly to f.

To prove the last statement, begin by noting that $\mathcal{F}_n f = f$ for any $f \in \mathbb{T}_n$; and moreover, note that the trigonometric polynomials are dense in $C_p(2\pi)$. It then follows from the Banach-Steinhaus theorem (Theorem 2.4.5) that there exist functions $f \in C_p(2\pi)$ for which $\mathcal{F}_n f$ does not converge uniformly to f. Note however that since such an f is in $L^2(-\pi, \pi)$, $\mathcal{F}_n f$ does converge to f in the L^2-norm.

In contrast to the above results in $C_p(2\pi)$, recall that $\mathcal{F}_n$ is an orthogonal projection operator with respect to $L^2(-\pi, \pi)$; and therefore $\|\mathcal{F}_n\| = 1$ when $\mathcal{F}_n$ is viewed as an operator from $L^(-\pi, \pi)$ to itself.

Uniform error bounds for the Fourier series

For a given function f and a given integer $n \geq 0$, let q_n denote the best approximation of f from the approximating subspace $\mathbb{T}_n$. Note that $\mathcal{F}_n(q_n) = q_n$. Then using the linearity of $\mathcal{F}_n$,

$$f - \mathcal{F}_n(f) = (f - q_n) - \mathcal{F}_n(f - q_n).$$

Taking norms of both sides,

$$\|f - \mathcal{F}_n(f)\|_\infty \leq (1 + \|\mathcal{F}_n\|)\, \|f - q_n\|_\infty .$$

Assuming $f \in C_p^{k,\alpha}(2\pi)$ for some $k \geq 0$ and some $\alpha \in (0, 1]$, we have

$$\|f - \mathcal{F}_n(f)\|_\infty \leq (1 + \|\mathcal{F}_n\|)\, c^{k+1} \frac{M_k}{n^{k+\alpha}} \tag{3.7.11}$$

and

$$\|f - \mathcal{F}_n(f)\|_\infty \leq c_k \frac{\log n}{n^{k+\alpha}} \quad \text{for } n \geq 2. \tag{3.7.12}$$

Here $c = 1 + \pi^2/2$, M_k is the Hölder constant for $f^{(k)}$ and c_k is a constant linearly dependent on M_k and otherwise independent of f. Combining this with (3.7.10), we see that if $f \in C_p^{0,\alpha}(2\pi)$ for some α, then $\mathcal{F}_n(f)$ converges uniformly to f. For $\mathcal{F}_n(f)$ to fail to converge uniformly to f, the function f must be fairly badly behaved.

3.7.2 L^2-approximations using polynomials

Recall the introduction to orthogonal polynomials in Section 3.5. Also, recall that the orthogonal projection operator from an inner product space V onto an approximation subspace V_N can be interpreted as a least squares approximation; see (3.4.6) in Section 3.4 and the derivation preceding it. As before in Section 3.5, consider the approximation of functions $u \in L_w^2(-1,1)$ based on the inner product

$$(u,v)_{0,w} = \int_{-1}^{1} u(x)\, v(x)\, w(x)\, dx,$$

introduced earlier in (3.5.1). The operator P_N of (3.5.2) is the orthogonal projection of $L_w^2(-1,1)$ onto $\mathbb{P}_N$, the space of polynomials of degree $\leq N$. In analogy with the earlier results for Fourier series, we want to analyze the uniform convergence of $P_N u$ to u when $u \in C[-1,1]$.

Let $\{p_n(x)\}_{n=0}^{\infty}$ denote orthogonal polynomials with respect to the above inner product $(\cdot,\cdot)_{0,w}$, with $\deg(p_n) = n$, $n \geq 0$. For notational simplicity and without loss of generality, assume $(p_n, p_n)_{0,w} = 1$, $n \geq 0$, thus making $\{p_n(x)\}_{n=0}^{\infty}$ an *orthonormal* basis for $L_w^2(-1,1)$. The set $\{p_n(x)\}_{n=0}^{N}$ is an orthonormal basis for $\mathbb{P}_N$: for $u \in L_w^2(-1,1)$,

$$P_N u(x) = \sum_{n=0}^{N} \xi_n p_n(x), \quad \xi_n = (u, p_n)_{0,w}, \quad 0 \leq n \leq N. \tag{3.7.13}$$

is the orthogonal projection of u onto $\mathbb{P}_N$. As with our analysis in §3.7.1 of the uniform convergence of Fourier series, we need to know $\|P_N\|$ when P_N is considered as an operator from $C[-1,1]$ to $\mathbb{P}_N \subset C[-1,1]$. In analogy with the derivation of (3.7.11), we can then obtain

$$\|u - P_N u\|_\infty \leq (1 + \|P_N\|)\, \|u - q\|_\infty, \tag{3.7.14}$$

where q is an arbitrary polynomial from $\mathbb{P}_N$. By choosing q to be the minimax approximation to u from $\mathbb{P}_N$ and by applying Jackson's Theorem 3.7.2, we obtain rates of convergence for the speed of uniform convergence of $P_N u$ to u.

To study the size of $\|P_N\|$, use (3.7.13) to obtain the integral formula

$$P_N u(x) = \int_{-1}^{1} K(x,t)\, u(t)\, dt, \quad -1 \le x \le 1, \tag{3.7.15}$$

$$K(x,t) = \sum_{n=0}^{N} p_n(x)\, p_n(t). \tag{3.7.16}$$

By regarding (3.7.15) as defining an integral operator from $C[-1,1]$ to itself, the earlier result (2.2.8) implies that

$$\|P_N\| = \max_{-1\le x\le 1} \int_{-1}^{1} |K(x,t)|\, dt. \tag{3.7.17}$$

The sum in (3.7.16) is complicated, but can be simplified using the following result.

Theorem 3.7.3 (Christoffel-Darboux Identity) *For $\{p_n(x)\}_{n=0}^{\infty}$ an orthonormal family of polynomials with weight function $w(x) \ge 0$,*

$$\sum_{n=0}^{N} p_n(x)\, p_n(t) = \begin{cases} \dfrac{p_{N+1}(x)\, p_N(t) - p_N(x)\, p_{N+1}(t)}{a_N (x-t)}, & x \ne t \\ \dfrac{p'_{N+1}(t)\, p_N(t) - p'_N(t)\, p_{N+1}(t)}{a_N}, & x = t \end{cases} \tag{3.7.18}$$

with $a_N = A_{N+1}/A_N$ and $p_n(x) = A_n x^n +$ lower order terms, $n \ge 0$.

For a proof, see [220, p. 43]. The proof is based on a manipulation of the triple recursion relation given in Exercises 3.5.5–3.5.6 of Section 3.5. The formula (3.7.18) leads immediately to a simpler expression for the kernel function of (3.7.15). This function $K(x,t)$ is an example of a *reproducing kernel function*, and this concept is introduced and explored in Exercise 3.2.10.

Example 3.7.4 Return to the Chebyshev polynomials $\{T_n(x)\}$ of (3.5.8). For this case, $A_n = 2^{n-1}$, $n \ge 1$, and $a_N = 2$, $N \ge 1$. Then (3.7.18) becomes

$$K(x,t) = \frac{\cos(N+1)\theta\, \cos N\phi - \cos N\theta\, \cos(N+1)\phi}{2\,(x-t)}, \quad x \ne t$$

with $x = \cos\theta$ and $t = \cos\phi$, $0 \le \theta, \phi \le \pi$. An alternative derivation, given in [195, p. 133], yields

$$K(x,t) = \frac{1}{2\pi}\left[\frac{\sin((N+1/2)\,(\theta+\phi))}{\sin((\theta+\phi)/2)} + \frac{\sin((N+1/2)\,(\theta-\phi))}{\sin((\theta-\phi)/2)}\right].$$

This can be used in (3.7.17) to bound $\|P_N\|$. Noting the periodicity of $K(x,t)$ in θ and ϕ, we can convert (3.7.17) to

$$\|P_N\| = \frac{1}{\pi}\int_0^\pi \frac{\sin[(2N+1)\,\phi/2]}{\sin(\phi/2)}\,d\phi.$$

Recall this same integral appeared in the error formula for the truncated Fourier series (cf. (3.7.8) and (3.7.9)). Combining the result (3.7.10) with the above, we obtain

$$\|P_N\| = \frac{4}{\pi^2}\log n + \mathcal{O}(1), \quad n \geq 1.$$

When this relation is combined with (3.7.14) and Jackson's Theorem 3.7.2, a bound for $\|u - P_N u\|_\infty$ can be obtained that looks exactly the same as that in (3.7.12) for Fourier series. This bound on $\|P_N\|$ and the resulting bound on $\|u - P_N u\|_\infty$ shows that the expansion of a function $u(x)$ in terms of Chebyshev polynomials converges uniformly for most functions $u \in C[-1,1]$.

3.7.3 Interpolatory projections and their convergence

Recall the trigonometric interpolation discussion of Subsection 3.2.4. Let $f \in C_p(2\pi)$, let $n \geq 0$ be a given integer, and let the interpolation nodes be the evenly spaced points in (3.2.16). Denote the resulting interpolation formula by $\mathcal{I}_n(f)$. It is straightforward to show that this is a linear operator; and by the uniqueness of such trigonometric polynomial interpolation, it also follows that $\mathcal{I}_n$ is a projection operator on $C_p(2\pi)$ to $\mathbb{T}_n$. To discuss the convergence of $\mathcal{I}_n(f)$ to f, we can proceed in the same manner as when examining the convergence of $\mathcal{F}_n(f)$.

Begin by obtaining the Lagrange interpolation formula

$$\mathcal{I}_n f(x) = \frac{2}{2n+1}\sum_{j=0}^{2n} D_n(x - x_j)\, f(x_j). \tag{3.7.19}$$

Its proof is left as Exercise 3.7.5. Using this formula,

$$\|\mathcal{I}_n\| = \frac{2}{2n+1}\max_x \sum_{j=0}^{2n} |D_n(x - x_j)|.$$

In Rivlin [195, p. 13], it is shown that

$$\|\mathcal{I}_n\| \leq 1 + \frac{2}{\pi}\log n, \quad n \geq 1 \tag{3.7.20}$$

and it is also shown that $\|\mathcal{I}_n\|$ is exactly of order $\log n$ as $n \to \infty$. This result can be combined with an argument such as the one leading to (3.7.12)

to obtain analogous results for the convergence of $\mathcal{I}_n f$. In fact, assuming $f \in C_p^{k,\alpha}(2\pi)$ for some $k \geq 0$ and some $\alpha \in (0, 1]$, we have

$$\|f - \mathcal{I}_n f\|_\infty \leq (1 + \|\mathcal{I}_n\|) c^{k+1} \frac{M_k}{n^{k+\alpha}} \tag{3.7.21}$$

for any $n \geq 1$, $c = 1 + \pi^2/2$, and

$$\|f - \mathcal{I}_n f\|_\infty \leq c_k \frac{\log n}{n^{k+\alpha}} \quad \text{for } n \geq 2 \tag{3.7.22}$$

for c_k a constant linearly dependent on M_k and otherwise independent of f.

Exercise 3.7.1 Show that $\cos(j\theta)$ can be written as $p_j(\cos\theta)$, with $p_j(x)$ a polynomial of degree j.

Exercise 3.7.2 Show that if $g(\theta)$ is an even 2π-periodic function, then its best approximation of degree n must take the form (3.7.2).

Exercise 3.7.3 Derive (3.7.6).

Exercise 3.7.4 Prove the formula (3.7.10).
Hint: Make the observation

$$\int_0^\pi |D_n(y)|\, dy = \int_0^\pi \left| \frac{\sin(n + 1/2)x}{x} \right| dx + \mathcal{O}(1) = \int_0^{n\pi} \frac{|\sin t|}{t}\, dt + \mathcal{O}(1).$$

Exercise 3.7.5 Show that the functions

$$\phi_j(x) \equiv \frac{2}{2n+1} D_n(x - x_j), \quad 0 \leq j \leq 2n,$$

belong to $\mathbb{T}_n$ and that they satisfy

$$\phi_j(x_k) = \delta_{jk}.$$

Thus show that (3.7.19) can be considered a "Lagrange interpolation formula".

Exercise 3.7.6 Show that $\mathcal{I}_n(f)$ can be obtained from $\mathcal{F}_n(f)$ by a suitably chosen numerical integration.

Suggestion for Further Reading.

Interpolation is a standard topic found in every textbook on numerical analysis. The reader is referred to Atkinson [15, Chap. 3], Kress [150] and other numerical analysis textbooks for a more detailed discussion of polynomial interpolation, as well as other interpolation topics not touched upon in this work, such as interpolation with spline functions. The classic

introduction to the theory of spline functions is De Boor [65]. Meinardus [169] is an excellent reference for approximations by polynomials and trigonometric polynomials. Davis [64] contains an extensive discussion of interpolation and approximation problems in a very general framework. Best approximations in inner product spaces are systematically treated in [69].

A best approximation problem is a minimization problem, with or without constraints, and some of them are best studied within the framework of *optimization theory*. Some abstract minimization problems are best studied in the framework of *convex analysis*, and some excellent references on convex analysis include Ekeland and Temam [76], Rockafellar [197] and Zeidler [247].

There is current work on generalizing to the multivariable case the use of polynomial approximation. For an extensive introduction to multivariable orthogonal polynomials, see Dunkl and Xu [72] and Xu [239]–[241]. For generalizations of the Jackson theorems to the unit ball in $\mathbb{R}^d$ and the unit sphere in $\mathbb{R}^{d+1}$, $d \geq 2$, see Ragozin [190].

4
Fourier Analysis and Wavelets

In this chapter, we provide an introduction to the theory of Fourier analysis and wavelets. Fourier analysis is a large branch of mathematics, and it is useful in a wide spectrum of applications, such as in solving differential equations arising in sciences and engineering, and in signal processing. The first three sections of the chapter will be devoted to the Fourier series, the Fourier transform, and the discrete Fourier transform, respectively. The Fourier transform converts a function of a time or space variable into a function of a frequency variable. When the original function is periodic, it is sufficient to consider integer multiples of the base frequency, and we are led to the notion of the Fourier series. For a general non-periodic function, we need coefficients of all possible frequencies, and the result is the Fourier transform.

The last two sections of the chapter are devoted to wavelets. Since the 1980s, the theory and applications of wavelets have become a major area of mathematical research. Wavelets and Fourier analysis complement each other in providing tools for efficient treatment of problems from a wide variety of areas in mathematics, the sciences, and engineering.

4.1 Fourier series

In solving several model partial differential equations with the method of separation of variables, a natural question is whether a function can be represented by a trigonometric series. Indeed, J. Fourier used extensively

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_4,
© Springer Science + Business Media, LLC 2009

the Fourier series in the study of the heat conduction, and he published his results in his famous work *Théorie Analytique de la Chaleur* in 1822.

Let $f \in L^1(-\pi, \pi)$. Then its Fourier series is defined by

$$F(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} \left[a_j \cos(jx) + b_j \sin(jx) \right], \tag{4.1.1}$$

where the Fourier coefficients are defined by

$$a_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(jx)\, dx, \quad j \geq 0, \tag{4.1.2}$$

$$b_j = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(jx)\, dx, \quad j \geq 1. \tag{4.1.3}$$

These formulas for the coefficients can be derived formally by multiplying both sides of (4.1.1) with $\cos(jx)$ or $\sin(jx)$, integrating over the interval $[-\pi, \pi]$, interchanging the order of integration and infinite summation, and replacing $F(x)$ by $f(x)$. The Fourier series $F(x)$ defined by (4.1.1)–(4.1.3) is closely related to the function $f(x)$; however, the series may not converge, and even when it converges, the limit may be different from $f(x)$. For this reason, we use the notation $F(x)$ for the Fourier series of $f(x)$. The convergence issue is examined later in this section.

For an integer $j > 0$, the harmonic modes $\cos(jx)$ and $\sin(jx)$ have the frequency (defined as the reciprocal of the period, i.e., the number of cycles per unit time) $\omega_j = j/(2\pi)$, which is the j-multiple of the base frequency $\omega_1 = 1/(2\pi)$. Thus, if the function $f(x)$ represents a periodic signal with period 2π, then the Fourier series (4.1.1) can be interpreted as a decomposition of the signal as a linear combination of a constant term and harmonic modes with frequencies $\{\omega_j\}_{j=1}^{\infty}$.

Since the function $F(x)$ of (4.1.1) is a 2π-periodic function, it is usually convenient to view $f \in L^1(-\pi, \pi)$ as being defined on $\mathbb{R}$ with period 2π. In the discussion of this section, we focus on the standard interval $[-\pi, \pi]$. For a function defined on a general interval, its Fourier series can be studied by relating the general interval to the standard one through a change of the independent variable; see Exercise 4.1.1.

Complex form. By the Euler identity

$$e^{i\theta} = \cos\theta + i \sin\theta,$$

we obtain

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2\,i}.$$

Using these formulas, we can rewrite (4.1.1) in the form

$$F(x) = \sum_{j=-\infty}^{\infty} c_j e^{ijx}, \tag{4.1.4}$$

where the Fourier coefficients

$$c_j = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\, e^{-ijx}dx, \quad -\infty < j < \infty. \tag{4.1.5}$$

This is called the complex form of the Fourier series of $f \in L^1(-\pi,\pi)$. The formula (4.1.1) gives the real form of the Fourier series. When f is real-valued, usually the real form (4.1.1) of the Fourier series is used, and the corresponding Fourier coefficients are real. Nevertheless, for any L^1 function, real or complex valued, both forms of the Fourier series can be used. Obviously, we have the relations

$$a_0 = 2\,c_0, \quad a_j = 2\,\Re(c_j), \quad b_j = -2\,\Im(c_j), \quad j = 1, 2, \cdots \tag{4.1.6}$$

between the Fourier coefficients defined in (4.1.2)–(4.1.3) and in (4.1.5). Here, $\Re(c_j)$ and $\Im(c_j)$ are the real and imaginary parts of c_j.

In the rest of the section, we will mainly consider the real form of the Fourier series; the theoretical results for the real form, such as convergence, are valid for the complex case also.

Sine and cosine series. Let $f \in L^1(-\pi,\pi)$ be an odd function, i.e., $f(-x) = -f(x)$ for $x \in [-\pi,\pi]$. Then its Fourier series reduces to a sine series (Exercise 4.1.2):

$$F(x) = \sum_{j=1}^{\infty} b_j \sin(jx), \tag{4.1.7}$$

where

$$b_j = \frac{2}{\pi}\int_0^{\pi} f(x)\, \sin(jx)\, dx, \quad j \geq 1. \tag{4.1.8}$$

Similarly, suppose $f \in L^1(-\pi,\pi)$ is an even function, i.e., $f(-x) = f(x)$ for $x \in [-\pi,\pi]$. Then its Fourier series reduces to a cosine series (Exercise 4.1.3):

$$F(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} a_j \cos(jx), \tag{4.1.9}$$

where

$$a_j = \frac{2}{\pi}\int_0^{\pi} f(x)\, \cos(jx)\, dx, \quad j \geq 0. \tag{4.1.10}$$

Given a function $f \in L^1(0,\pi)$, we can develop a sine series for it. This is achieved as follows. First, we extend $f(x)$ to an odd function on $[-\pi,\pi]$:

$$\tilde{f}(x) = \begin{cases} f(x), & 0 \leq x \leq \pi, \\ -f(-x), & -\pi \leq x < 0. \end{cases}$$

Strictly speaking, $\tilde{f}$ is an odd function only if $f(0) = 0$. Nevertheless, even without this property, the Fourier series of $\tilde{f}$ is a sine series since the

coefficients of the Fourier series are computed from integrals and do not depend on the function value at any particular point. Then, we use the sine series of $\tilde{f}$ to be that for f:

$$F(x) = \sum_{j=1}^{\infty} b_j \sin(jx)$$

with

$$b_j = \frac{2}{\pi} \int_0^{\pi} f(x) \sin(jx)\, dx, \quad j \geq 1.$$

We can develop a cosine series for the function $f \in L^1(0, \pi)$ as well. First, we extend f to an even function on $[-\pi, \pi]$:

$$\tilde{f}(x) = \begin{cases} f(x), & 0 \leq x \leq \pi, \\ f(-x), & -\pi \leq x \leq 0. \end{cases}$$

Then we use the cosine series of $\tilde{f}$ for f:

$$F(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} a_j \cos(jx)$$

with

$$a_j = \frac{2}{\pi} \int_0^{\pi} f(x) \cos(jx)\, dx, \quad j \geq 0.$$

Convergence. We now turn to a discussion of convergence of the Fourier series in various sense. Pointwise convergence of the Fourier series (4.1.1) to the function $f(x)$ is delicate to analyze. We have the following result.

Theorem 4.1.1 *Assume f is a piecewise continuous, 2π-periodic function. Let $x \in [-\pi, \pi]$ (or $x \in \mathbb{R}$ due to the periodicity of f and its Fourier series F) be a point where the two one-sided derivatives $f'(x-)$ and $f'(x+)$ exist. Then*

$$F(x) = \frac{1}{2} \left[f(x-) + f(x+) \right].$$

In particular, if f is continuous at x, then

$$F(x) = f(x),$$

i.e.,

$$f(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} \left[a_j \cos(jx) + b_j \sin(jx) \right]$$

where $\{a_j\}$ and $\{b_j\}$ are defined in (4.1.2) *and* (4.1.3).

A proof of this theorem is the subject of Exercise 4.1.4. A similar result holds for the complex form of the Fourier series. We emphasize that even when the Fourier series (4.1.1) for a continuous function is convergent at a point, the limit does not need to be the value of the function at the point.

From the discussion of Example 1.3.15, we always have convergence of the Fourier series in L^2 norm: for $f \in L^2(-\pi, \pi)$,

$$\int_{-\pi}^{\pi} \left\{ f(x) - \left[\frac{a_0}{2} + \sum_{j=1}^{n} (a_j \cos(jx) + b_j \sin(jx)) \right] \right\}^2 dx \to 0 \quad \text{as } n \to \infty.$$

We simply write

$$f(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} [a_j \cos(jx) + b_j \sin(jx)] \quad \text{in } L^2(-\pi, \pi).$$

This implies

$$f(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} [a_j \cos(jx) + b_j \sin(jx)] \quad \text{a.e. } x \in [-\pi, \pi].$$

Turning now to the issue of convergence in a general L^p-norm, we define the partial sum sequence

$$S_N f(x) = \frac{a_0}{2} + \sum_{j=1}^{N} [a_j \cos(jx) + b_j \sin(jx)]$$

with the coefficients $a_0, a_1, \ldots, a_N$ and $b_1, \ldots, b_N$ given by (4.1.2) and (4.1.3). We have the following result.

Theorem 4.1.2 *Let $1 \leq p < \infty$. Then*

$$\|S_N f - f\|_{L^p(-\pi,\pi)} \to 0 \quad \text{as } N \to \infty, \quad \forall\, f \in L^p(-\pi, \pi) \tag{4.1.11}$$

if and only if there exists a constant $C_p > 0$ such that

$$\|S_N f\|_{L^p(-\pi,\pi)} \leq C_p \|f\|_{L^p(-\pi,\pi)} \quad \forall\, N \geq 1,\ \forall\, f \in L^p(-\pi, \pi). \tag{4.1.12}$$

Proof. For each positive integer N, the operator S_N is obviously linear and continuous on $L^p(-\pi, \pi)$. Assume (4.1.11) is true. Then for any $f \in L^p(-\pi, \pi)$, the sequence $\{\|S_N f\|_{L^p(-\pi,\pi)}\}$ is bounded. By the principle of uniform boundedness (Theorem 2.4.4), we know that the operator sequence $\{S_N\} \subset \mathcal{L}(L^p(-\pi, \pi))$ is bounded.

Now assume (4.1.12). Let $f \in L^p(-\pi, \pi)$ and $\epsilon > 0$. We use the density of the trigonometric polynomials in $L^p(-\pi, \pi)$. Select a trigonometric polynomial f_ϵ such that

$$\|f - f_\epsilon\|_{L^p(-\pi,\pi)} < \epsilon.$$

For $N \geq \deg f_\epsilon$, we have $S_N f_\epsilon = f_\epsilon$. Hence,

$$S_N f - f = S_N(f - f_\epsilon) + (f_\epsilon - f)$$

and

$$\begin{aligned} \|S_N f - f\|_{L^p(-\pi,\pi)} &\leq \|S_N(f - f_\epsilon)\|_{L^p(-\pi,\pi)} + \|f - f_\epsilon\|_{L^p(-\pi,\pi)} \\ &\leq (C_p + 1)\,\epsilon. \end{aligned}$$

Therefore, we have the convergence (4.1.11). □

It can be shown that for $1 < p < \infty$, (4.1.12) holds, and so we have the convergence of the Fourier series in $L^p(-\pi,\pi)$ for any $L^p(-\pi,\pi)$ function (see [73, Chapter 3]). In particular, it is easy to verify (4.1.12) in the case $p = 2$ (Exercise 4.1.6). On the other hand, (4.1.12) does not hold for $p = 1$. Note that a consequence of the L^2-norm convergence of the Fourier series of $f \in L^2(-\pi,\pi)$ is the Parseval equality (cf. (1.3.10)):

$$\|f\|^2_{L^2(-\pi,\pi)} = \pi \left[\frac{|a_0|^2}{2} + \sum_{j=1}^{\infty} \left(|a_j|^2 + |b_j|^2 \right) \right]. \tag{4.1.13}$$

We now present several examples, showing the dependence of the convergence behavior of the Fourier series on smoothness of the function.

Example 4.1.3 Consider a step function

$$f(x) = \begin{cases} 1, & -\pi \leq x < 0, \\ 0, & 0 \leq x < \pi. \end{cases}$$

The Fourier coefficients of the function are

$$\begin{aligned} &a_0 = 1, \\ &a_j = 0, \quad j \geq 1, \\ &b_j = -\frac{1}{\pi j}\left[1 - (-1)^j\right], \quad j \geq 1. \end{aligned}$$

Thus, the Fourier series of the step function f is

$$F(x) = \frac{1}{2} - \frac{2}{\pi} \sum_{j=1}^{\infty} \frac{\sin(2j-1)x}{2j-1}.$$

By Theorem 4.1.1, we have the following pointwise convergence property (f is viewed as extended to outside $[-\pi,\pi]$ by period 2π):

$$F(x) = \begin{cases} f(x), & x \neq k\pi,\ k \text{ integer}, \\ 1/2, & x = k\pi,\ k \text{ integer}. \end{cases}$$

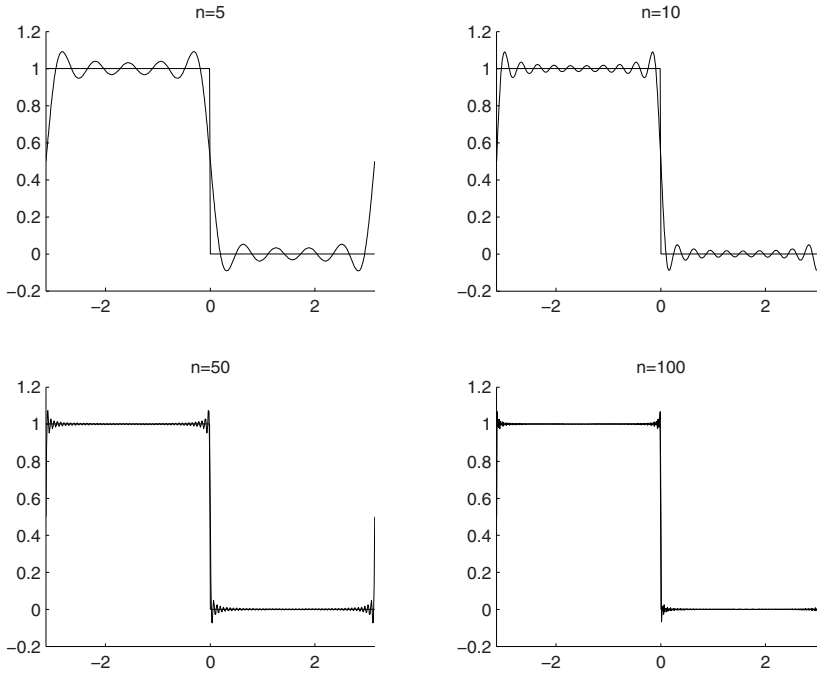

FIGURE 4.1. $f(x)$ and $S_n(x)$: Gibbs phenomenon

Figure 4.1 shows the step function $f(x)$ and partial sums

$$S_n(x) = \frac{1}{2} - \frac{2}{\pi} \sum_{j=1}^{n} \frac{\sin(2j-1)x}{2j-1}$$

with $n = 5, 10, 50$ and 100, for $x \in [-\pi, \pi]$. We clearly observe the convergence $S_n(x) \to f(x)$ for any fixed $x \in (-\pi, 0) \cup (0, \pi)$. Let us pay close attention to the convergence behavior of the partial sums $S_n(x)$ around the discontinuity point $x = 0$. We see that the partial sums $S_n(x)$ overshoot the function value at the upper side of the discontinuity and undershoot the function value at the lower side. As the number n increases, the accuracy of the approximation by $S_n(x)$ to $f(x)$ increases, and ripples move closer toward the discontinuity. However, the size of the ripples does not decrease to zero. This phenomenon is common for the Fourier series of a general function at a discontinuity point, and is termed Gibbs phenomenon. Let x be a discontinuity point of a general function f. For the partial sums of the Fourier series of such a function, it can be shown that the vertical span extending from the top of the overshoot to the bottom of the undershoot

approaches the value

$$\frac{2}{\pi}\,\mathrm{Sint}(\pi)\,|f(x+)-f(x-)|\,,$$

where

$$\mathrm{Sint}(x)=\int_0^x \frac{\sin t}{t}\,dt$$

is the sine integral function. We have

$$\frac{2}{\pi}\,\mathrm{Sint}(\pi)\approx 1.17898.$$

The reader is referred to [123] for a review of the theory and history of the Gibbs phenomenon. □

Example 4.1.4 In the second example, let

$$f(x)=\frac{|x|}{\pi},\quad -\pi\le x\le\pi.$$

It is easy to see $f(-\pi)=f(\pi)$, and this implies the continuity of the periodic extension of the function, again denoted as $f(x)$:

$$f(x)=\frac{|x-2k\pi|}{\pi},\quad (2k-1)\,\pi\le x\le(2k+1)\,\pi,\ k\in\mathbb{Z}.$$

Since the function is even, its Fourier series does not contain the sine terms. After some calculation of the coefficients, we obtain the Fourier series of the given function:

$$F(x)=\frac{1}{2}-\frac{4}{\pi^2}\sum_{j=1}^{\infty}\frac{\cos(2j-1)x}{(2j-1)^2}.$$

We have $F(x)=f(x)$ both pointwise and in $L^2(-\pi,\pi)$. For any integer $n\ge 1$, define the partial sum

$$S_n(x)=\frac{1}{2}-\frac{4}{\pi^2}\sum_{j=1}^{n}\frac{\cos(2j-1)x}{(2j-1)^2}.$$

In Figure 4.2, we plot $f(x)$ and the Fourier series partial sum $S_2(x)$. The difference between the two curves is visible, but the error $[f(x)-S_2(x)]$ is not big. We then increase the number of terms for the partial sum. Figure 4.3 shows graphs of $f(x)$ and the Fourier series partial sum $S_{10}(x)$. We see that the difference between the two functions is almost invisible, indicating a good accuracy of the approximation $S_{10}(x)$. □

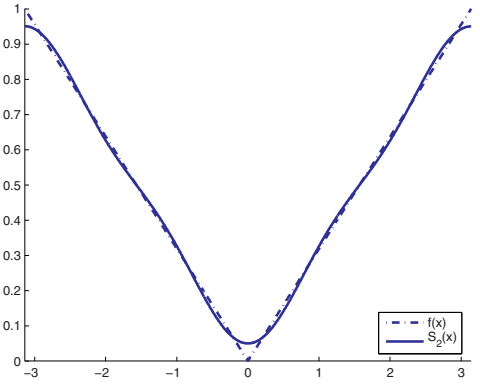

FIGURE 4.2. $f(x)$ and its Fourier series partial sum $S_2(x)$ for Example 4.1.4

Example 4.1.5 In this last example, we consider a smoother function

$$f(x) = \frac{(\pi^2 - x^2)^2}{\pi^4}, \quad -\pi \le x \le \pi.$$

We can verify that

$$f^{(l)}(-\pi) = f^{(l)}(\pi), \quad l = 0, 1, 2.$$

Thus, the periodic extension of the function, again denoted as $f(x)$, is twice continuously differentiable. The extended function is

$$f(x) = \frac{\left[\pi^2 - (x - 2k\pi)^2\right]^2}{\pi^4}, \quad (2k-1)\pi \le x \le (2k+1)\pi, \ k \in \mathbb{Z}.$$

This function is equal to its Fourier series pointwise:

$$f(x) = \frac{8}{15} + \frac{48}{\pi^4} \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{j^4} \cos(jx).$$

For an integer $n \ge 1$, define the partial sum

$$S_n(x) = \frac{8}{15} + \frac{48}{\pi^4} \sum_{j=1}^{n} \frac{(-1)^{j+1}}{j^4} \cos(jx).$$

Figure 4.4 shows the function $f(x)$ and its Fourier series partial sum $S_2(x)$. We observe a very good accuracy in approximating $f(x)$ by $S_2(x)$, even though only 3 terms are used in the approximation. □

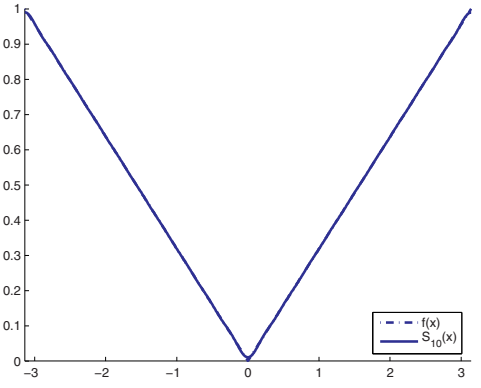

FIGURE 4.3. $f(x)$ and its Fourier series partial sum $S_{10}(x)$ for Example 4.1.4

From these three examples, we see clearly that the convergence behavior of the Fourier series depends on the smoothness of the (periodically extended) function. The smoothness of the extended function is equivalent to the smoothness of the function on the given interval $[-\pi, \pi]$ and the degree of the periodicity of the function described by the end-point conditions

$$f^{(l)}(\pi) = f^{(l)}(-\pi), \quad l = 0, 1, \cdots, k-2.$$

This condition does not hold for the function in Example 4.1.3, and is satisfied with $k = 2$ and 4 for the functions in Example 4.1.4 and Example 4.1.5, respectively. In general, with the Fourier series (4.1.1) for an $L^2(-\pi, \pi)$ function $f(x)$, if we define the partial sums

$$S_n(x) = \frac{a_0}{2} + \sum_{j=1}^{n} [a_j \cos(jx) + b_j \sin(jx)]$$

for positive integers n, then by the Parseval's equality (4.1.13) we have

$$\begin{aligned} \|f - S_n\|^2_{L^2(-\pi,\pi)} &= \left\| \sum_{j=n+1}^{\infty} [a_j \cos(jx) + b_j \sin(jx)] \right\|^2_{L^2(-\pi,\pi)} \\ &= \pi \sum_{j=n+1}^{\infty} \left(|a_j|^2 + |b_j|^2 \right). \end{aligned} \tag{4.1.14}$$

When

$$\sum_{j=1}^{\infty} (|a_j| + |b_j|) \leq \infty,$$

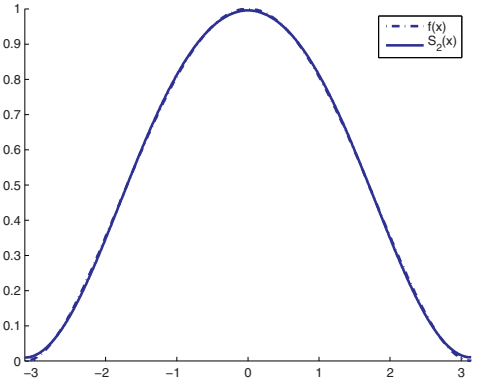

FIGURE 4.4. $f(x)$ and its Fourier series partial sum $S_2(x)$ for Example 4.1.5

the Fourier series (4.1.1) is absolutely convergent, and is thus continuous. In this case, the function f is continuous after possibly being modified on a set of measure zero. Then for the approximation error

$$f(x) - S_n(x) = \sum_{j=n+1}^{\infty} \left[a_j \cos(jx) + b_j \sin(jx) \right],$$

we have the bound

$$\max_{-\pi \le x \le \pi} |f(x) - S_n(x)| \le \sum_{j=n+1}^{\infty} \left(|a_j| + |b_j| \right). \tag{4.1.15}$$

Both (4.1.14) and (4.1.15) imply that the convergence speed of the partial sums of the Fourier series is closely related to the decay rate of the Fourier coefficients. Exercise 4.1.7 derives decay rate of the Fourier coefficients in relation to the smoothness and the degree of the periodicity of the given function.

Exercise 4.1.1 Show that the Fourier series of $f \in L^1(x_0 - T/2, x_0 + T/2)$ (T can be interpreted as the period of f viewed as being defined on $\mathbb{R}$) is

$$F(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} \left[a_j \cos(2j\pi(x - x_0)/T) + b_j \sin(2j\pi(x - x_0)/T) \right],$$

where

$$a_j = \frac{2}{T} \int_{x_0 - T/2}^{x_0 + T/2} f(x) \, \cos(2j\pi(x - x_0)/T) \, dx, \quad j \ge 0,$$

$$b_j = \frac{2}{T} \int_{x_0 - T/2}^{x_0 + T/2} f(x) \, \sin(2j\pi(x - x_0)/T) \, dx, \quad j \ge 1.$$

Derive these formulas from (4.1.1)–(4.1.3) with the change of variable: $x = x_0 + Tt/(2\pi)$ for $t \in [-\pi, \pi]$ and $x \in [x_0 - T/2, x_0 + T/2]$.

Exercise 4.1.2 Show that the Fourier series of an odd function $f \in L^1(-\pi, \pi)$ is given by (4.1.7)–(4.1.8).

Exercise 4.1.3 Show that the Fourier series of an even function $f \in L^1(-\pi, \pi)$ is given by (4.1.9)–(4.1.10).

Exercise 4.1.4 In this exercise, we present a proof of Theorem 4.1.1. Carry out the proof in the following steps.
(a) For any positive integer n, consider the partial sum

$$S_n f(x) = \frac{a_0}{2} + \sum_{j=1}^{n} \left[a_j \cos(jx) + b_j \sin(jx) \right].$$

Using the formulas for the coefficients a_j and b_j, show that (cf. (3.7.6))

$$S_n f(x) = \int_{-\pi}^{\pi} f(t)\, K_n(t - x)\, dt,$$

where

$$K_n(t) = \frac{1}{\pi} \left[\frac{1}{2} + \sum_{j=1}^{n} \cos(jt) \right].$$

(b) Regarding the kernel function K_n, prove the formulas

$$\int_{-\pi}^{0} K_n(t)\, dt = \int_{0}^{\pi} K_n(t)\, dt = \frac{1}{2}$$

and

$$K_n(t) = \begin{cases} \dfrac{\sin((n + 1/2)\, t)}{2\, \pi\, \sin(t/2)}, & t \neq 2k\pi,\ k \text{ integer}, \\ \dfrac{n + 1/2}{\pi}, & t = 2k\pi,\ k \text{ integer}. \end{cases}$$

(c) By a change of variables, write the partial sum as

$$S_n f(x) = \int_{-\pi - x}^{\pi - x} f(x + t)\, K_n(t)\, dt,$$

which can be rewritten as

$$S_n f(x) = \int_{-\pi}^{\pi} f(x + t)\, K_n(t)\, dt$$

due to the periodicity of both f and K_n.
(d) Using the results from (b) and (c), write

$$S_n f(x) - \frac{1}{2} \left[f(x-) + f(x+) \right] = \int_{0}^{\pi} \left[f(x + t) - f(x+) \right] K_n(t)\, dt + \int_{-\pi}^{0} \left[f(x + t) - f(x-) \right] K_n(t)\, dt.$$

Consider the first integral, written as

$$I_{n,+} = \int_0^\pi g_+(t)\, \sin((n+1/2)\, t)\, dt,$$

where

$$g_+(t) = \begin{cases} \dfrac{f(x+t) - f(x+)}{2\,\pi\, \sin(t/2)}, & t \in (0, \pi], \\ \dfrac{1}{\pi}\, f'(x+), & t = 0. \end{cases}$$

Observe that g_+ is piecewise continuous on $(0, \pi]$ and is right continuous at $t = 0$. By the Riemann-Lebesgue lemma (see Exercise 4.1.5),

$$I_{n,+} \to 0 \quad \text{as } n \to \infty.$$

Similarly, the second integral can be shown to converge to zero as $n \to \infty$. Hence, taking the limit $n \to \infty$, we obtain

$$F(x) = \frac{1}{2}\left[f(x-) + f(x+)\right].$$

Exercise 4.1.5 Riemann-Lebesgue Lemma: If $f \in L^1(a, b)$, then

$$\int_a^b f(t)\, \sin(st)\, dt \to 0 \quad \text{as } s \to \infty.$$

The result holds also with $\cos(st)$ replacing $\sin(st)$.

A proof of this result using the density of $C^1[a, b]$ in $L^1(a, b)$ is as follows: for any $\epsilon > 0$, let $f_\epsilon \in C^1[a, b]$ such that

$$\|f - f_\epsilon\|_{L^1(a,b)} < \epsilon/2.$$

Then for the fixed ϵ, show that

$$\left| \int_a^b f_\epsilon(t)\, \sin(st)\, dt \right| < \frac{\epsilon}{2}$$

for s large enough. Provide the detailed argument.

Also, give an elementary proof of the result by using the density of step functions (i.e. piecewise constant functions) in $L^1(a, b)$.

Exercise 4.1.6 Prove (4.1.12) for the case $p = 2$.
Hint: Apply the Parseval's equality (1.3.7).

Exercise 4.1.7 Convergence and convergence speed of the Fourier series (4.1.1) or (4.1.4) are determined by the decay behavior of the Fourier coefficients.
(a) Assume f is k times continuously differentiable and $f^{(k+1)} \in L^1(-\pi, \pi)$. Show that for $j \neq 0$, the coefficient c_j defined in (4.1.5) has the expansion

$$c_j = \frac{1}{2\pi} \sum_{l=0}^{k} (-1)^{l+j} \left(\frac{i}{j}\right)^{l+1} \left[f^{(l)}(\pi) - f^{(l)}(-\pi)\right] + \frac{1}{2\pi} \left(-\frac{i}{j}\right)^{k+1} \int_{-\pi}^{\pi} f^{(k+1)}(x)\, e^{-ijx} dx.$$

(b) Use the result from (a) and the relation (4.1.6) to show the following. Assume $f^{(2k-1)}$ is continuous and $f^{(2k)} \in L^1(-\pi,\pi)$. Then for $j = 1, 2, \cdots$,

$$\begin{aligned} a_j &= \frac{1}{\pi} \sum_{l=0}^{k-1} \frac{(-1)^{j+l}}{j^{2l+2}} \left[f^{(2l+1)}(\pi) - f^{(2l+1)}(-\pi) \right] \\ &\quad + \frac{(-1)^k}{\pi j^{2k}} \int_{-\pi}^{\pi} f^{(2k)}(x) \cos(jx)\, dx, \\ b_j &= \frac{1}{\pi} \sum_{l=0}^{k-1} \frac{(-1)^{j+l+1}}{j^{2l+1}} \left[f^{(2l)}(\pi) - f^{(2l)}(-\pi) \right] \\ &\quad + \frac{(-1)^k}{\pi j^{2k}} \int_{-\pi}^{\pi} f^{(2k)}(x) \sin(jx)\, dx. \end{aligned}$$

Assume $f^{(2k)}$ is continuous and $f^{(2k+1)} \in L^1(-\pi,\pi)$. Then for $j = 1, 2, \ldots$, integrate by parts on the two integral terms above to obtain

$$\begin{aligned} a_j &= \frac{1}{\pi} \sum_{l=0}^{k-1} \frac{(-1)^{j+l}}{j^{2l+2}} \left[f^{(2l+1)}(\pi) - f^{(2l+1)}(-\pi) \right] \\ &\quad + \frac{(-1)^{k+1}}{\pi j^{2k+1}} \int_{-\pi}^{\pi} f^{(2k+1)}(x) \sin(jx)\, dx, \\ b_j &= \frac{1}{\pi} \sum_{l=0}^{k} \frac{(-1)^{j+l+1}}{j^{2l+1}} \left[f^{(2l)}(\pi) - f^{(2l)}(-\pi) \right] \\ &\quad + \frac{(-1)^k}{\pi j^{2k+1}} \int_{-\pi}^{\pi} f^{(2k+1)}(x) \cos(jx)\, dx. \end{aligned}$$

(c) Show that if for some integer $k > 0$, f satisfies

$$f^{(l)}(\pi) = f^{(l)}(-\pi), \quad l = 0, 1, \cdots, k-2$$

and $f^{(k)} \in L^1(-\pi,\pi)$, then the following bounds hold:

$$|a_j| + |b_j| \le c_0 j^{-k}, \quad j = 1, 2, \ldots$$

for some constant c_0 depending on f.

Exercise 4.1.8 Assume $f \in C_p^{m-1}(2\pi)$ for some integer $m \ge 1$, and assume $f^{(m)} \in L^q(-\pi,\pi)$ for some $q \in [1,\infty]$. In the Fourier series of (4.1.1), show that the coefficients $\{a_j, b_j\}$ satisfy

$$|a_j|, |b_j| \le \frac{c}{j^m} \|f^{(m)}\|_{L^q(-\pi,\pi)}, \quad j \ge 1$$

with c dependent on only q.
Hint: Integrate by parts.

4.2 Fourier transform

The Fourier transform can be viewed as a continuous form of the Fourier series. To introduce the Fourier transform, we consider the Fourier series of a function on the interval $[-L, L]$ and let $L \to \infty$. More precisely, let f be a smooth function with period $2L$. Then by the pointwise convergence theorem (the complex version), we have

$$f(x) = \sum_{j=-\infty}^{\infty} c_j e^{ij\pi x/L},$$

where

$$c_j = \frac{1}{2L} \int_{-L}^{L} f(t)\, e^{-ij\pi t/L} dt, \quad -\infty < j < \infty.$$

Thus,

$$f(x) = \sum_{j=-\infty}^{\infty} \left[\frac{1}{2L} \int_{-L}^{L} f(t)\, e^{-ij\pi t/L} dt \right] e^{ij\pi x/L}.$$

Let $\xi_j = j\pi/L$, $\Delta\xi = \pi/L$, and define

$$F_L(\xi) = \frac{1}{2\pi} \int_{-L}^{L} f(t)\, e^{-i\xi t} dt.$$

Then

$$f(x) = \sum_{j=-\infty}^{\infty} F_L(\xi_j)\, e^{i\xi_j x} \Delta\xi.$$

For large L, this summation can be viewed as a Riemann sum. Taking the limit $L \to \infty$ and noting that $F_L(\xi)$ formally approaches

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, e^{-i\xi t} dt,$$

we would expect the identity

$$\begin{aligned} f(x) &= \int_{-\infty}^{\infty} \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} f(t)\, e^{-i\xi t} dt \right] e^{i\xi x} d\xi \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)\, e^{-i\xi t} dt \right] e^{i\xi x} d\xi. \end{aligned}$$

It is thus natural to define the Fourier transform of f to be

$$F(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-i\xi x} dx, \tag{4.2.1}$$

and to expect the following Fourier inversion formula

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\xi)\, e^{i\xi x} d\xi. \tag{4.2.2}$$

We will treat the Fourier transform over the general d-dimensional space $\mathbb{R}^d$. Extending (4.2.1) to the multi-variable case, we use the formula

$$\mathcal{F}(f)(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(\boldsymbol{x})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, dx \tag{4.2.3}$$

to define the Fourier transform of $f \in L^1(\mathbb{R}^d)$. For convenience, we will usually use $\hat{f}$ to denote the Fourier transform of f:

$$\hat{f} \equiv \mathcal{F}(f).$$

It is easily seen that $\mathcal{F}$ is linear and bounded from $L^1(\mathbb{R}^d)$ to $L^\infty(\mathbb{R}^d)$:

$$\begin{aligned} &\mathcal{F}(\alpha\, f + \beta\, g) = \alpha\, \mathcal{F}(f) + \beta\, \mathcal{F}(g) \quad \forall\, f, g \in L^1(\mathbb{R}^d),\ \alpha, \beta \in \mathbb{C}, \\ &\|\mathcal{F}(f)\|_{L^\infty(\mathbb{R}^d)} \le (2\pi)^{-d/2} \|f\|_{L^1(\mathbb{R}^d)} \quad \forall\, f \in L^1(\mathbb{R}^d). \end{aligned}$$

Applying the Lebesgue dominated convergence theorem, Theorem 1.2.26, we see that $\mathcal{F}(f) \in C(\mathbb{R}^d)$. Moreover, by Riemann-Lebesgue Lemma (the one dimensional case is discussed in Exercise 4.1.5), we have

$$\hat{f}(\boldsymbol{\xi}) \to 0 \quad \text{as } \|\boldsymbol{\xi}\| \to \infty.$$

Other properties of the Fourier transform can be proved similarly using tools from the subject of Real Analysis (see [212, Chapter 1, Section 1]). In particular, when $\hat{f} \in L^1(\mathbb{R}^d)$, the Fourier inversion formula holds:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi})\, e^{i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\xi \quad \text{a.e. } \boldsymbol{x} \in \mathbb{R}^d. \tag{4.2.4}$$

The next step in the development of the theory would be to extend the definition of the Fourier transform from $L^1(\mathbb{R}^d)$ to $L^2(\mathbb{R}^d)$. Such an extension is achieved by a density argument based on the density of the space $L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ in $L^2(\mathbb{R}^d)$ and the identity

$$\|\hat{f}\|_{L^2(\mathbb{R}^d)} = \|f\|_{L^2(\mathbb{R}^d)} \quad \forall\, f \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d).$$

A better approach is through the theory of tempered distributions; see e.g., [73, 210]. Here, we sketch the main ideas of the theory.

Definition 4.2.1 *The space of test functions of rapid decay, known as the Schwartz space, $\mathcal{S}(\mathbb{R}^d)$, consists of smooth functions $\phi \in C^\infty(\mathbb{R}^d)$ such that for any multi-indices α and β,*

$$\boldsymbol{x}^\beta \partial^\alpha \phi(\boldsymbol{x}) \to 0 \quad \text{as } \|\boldsymbol{x}\| \to \infty.$$

Given $\{\phi, \phi_1, \phi_2, \cdots\} \subset \mathcal{S}(\mathbb{R}^d)$*, we say* ϕ_n *converges to* ϕ *in* $\mathcal{S}(\mathbb{R}^d)$ *if for any multi-indices* α *and* β*,*

$$\lim_{n\to\infty} \max_{\boldsymbol{x}\in\mathbb{R}^d} \left|\boldsymbol{x}^\beta \partial^\alpha [\phi(\boldsymbol{x}) - \phi_n(\boldsymbol{x})]\right| = 0.$$

Recall that $C_0^\infty(\mathbb{R}^d)$ denotes the space of all functions from $C^\infty(\mathbb{R}^d)$ that have compact support. Notice that algebraically, $C_0^\infty(\mathbb{R}^d) \subset \mathcal{S}(\mathbb{R}^d)$, but not conversely. For example, the function $\exp(-\|\boldsymbol{x}\|^2)$ belongs to $\mathcal{S}(\mathbb{R}^d)$ but not to $C_0^\infty(\mathbb{R}^d)$.

For any $f \in \mathcal{S}(\mathbb{R}^d)$, we use the formula (4.2.3) to define its Fourier transform. We list below some properties of the Fourier transform:

$$\mathcal{F}(\alpha f + \beta g) = \alpha \mathcal{F}(f) + \beta \mathcal{F}(g); \tag{4.2.5}$$

$$\|\mathcal{F}(f)\|_{L^\infty(\mathbb{R}^d)} \le (2\pi)^{-d/2} \|f\|_{L^1(\mathbb{R}^d)}; \tag{4.2.6}$$

$$\mathcal{F}(\partial^\alpha f)(\boldsymbol{\xi}) = (i\,\boldsymbol{\xi})^\alpha \mathcal{F}(f)(\boldsymbol{\xi}); \tag{4.2.7}$$

$$\mathcal{F}(\boldsymbol{x}^\alpha f(\boldsymbol{x})) = i^{|\alpha|} \partial^\alpha \mathcal{F}(f(\boldsymbol{x})). \tag{4.2.8}$$

These properties are quite easy to prove for functions in $\mathcal{S}(\mathbb{R}^d)$ (Exercise 4.2.1). We now present three more properties that are crucial for the extension of the definition of the Fourier transform.

$$\mathcal{F} \text{ is continuous from } \mathcal{S}(\mathbb{R}^d) \text{ to } \mathcal{S}(\mathbb{R}^d); \tag{4.2.9}$$

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\, \hat{g}(\boldsymbol{x})\, dx = \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{x})\, g(\boldsymbol{x})\, dx \quad \forall\, f, g \in \mathcal{S}(\mathbb{R}^d); \tag{4.2.10}$$

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi})\, e^{i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\xi \quad \forall\, f \in \mathcal{S}(\mathbb{R}^d). \tag{4.2.11}$$

The proof of (4.2.9) is routine, and is left as Exercise 4.2.3. The identity (4.2.10) is proved by an application of the Fubini theorem on the function $f(\boldsymbol{x})\, g(\boldsymbol{y}) \in L^1(\mathbb{R}^d \times \mathbb{R}^d)$:

$$\begin{aligned}
\int_{\mathbb{R}^d} f(\boldsymbol{x})\, \hat{g}(\boldsymbol{x})\, dx &= \int_{\mathbb{R}^d} f(\boldsymbol{x})\, (2\pi)^{-d/2} \int_{\mathbb{R}^d} g(\boldsymbol{y})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}\, dy\, dx \\
&= (2\pi)^{-d/2} \int_{\mathbb{R}^d\times\mathbb{R}^d} f(\boldsymbol{x})\, g(\boldsymbol{y})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}\, dx\, dy \\
&= \int_{\mathbb{R}^d} \left[(2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\boldsymbol{x})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}\, dx\right] g(\boldsymbol{y})\, dy \\
&= \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{x})\, g(\boldsymbol{x})\, dx.
\end{aligned}$$

Finally, we prove the inversion formula (4.2.11). Let $\boldsymbol{x}_0 \in \mathbb{R}^d$ be fixed and $\lambda > 0$ be a parameter. Denote

$$f_{\boldsymbol{x}_0,\lambda}(\boldsymbol{x}) = f(\boldsymbol{x}_0 + \lambda^{-1}\boldsymbol{x}).$$

Its Fourier transform is calculated as follows:

$$\mathcal{F}(f_{\boldsymbol{x}_0,\lambda})(\boldsymbol{y}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\boldsymbol{x}_0 + \lambda^{-1}\boldsymbol{x})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{y}}\, dx.$$

Introduce the change of the variable $\boldsymbol{z} = \boldsymbol{x}_0 + \lambda^{-1}\boldsymbol{x}$,

$$\begin{aligned}\mathcal{F}(f_{\boldsymbol{x}_0,\lambda})(\boldsymbol{y}) &= e^{i\lambda\boldsymbol{x}_0\cdot\boldsymbol{y}}\lambda^d(2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\boldsymbol{z})\, e^{-i\boldsymbol{z}\cdot\lambda\boldsymbol{y}}\, dz \\ &= e^{i\lambda\boldsymbol{x}_0\cdot\boldsymbol{y}}\lambda^d \hat{f}(\lambda\boldsymbol{y}).\end{aligned}$$

By (4.2.10),

$$\int_{\mathbb{R}^d} f(\boldsymbol{x}_0 + \lambda^{-1}\boldsymbol{x})\, \hat{g}(\boldsymbol{x})\, dx = \int_{\mathbb{R}^d} e^{i\lambda\boldsymbol{x}_0\cdot\boldsymbol{y}}\lambda^d \hat{f}(\lambda\boldsymbol{y})\, g(\boldsymbol{y})\, dy.$$

With a change of the variable $\boldsymbol{\xi} = \lambda\boldsymbol{y}$ for the integral on the right side, we have

$$\int_{\mathbb{R}^d} f(\boldsymbol{x}_0 + \lambda^{-1}\boldsymbol{x})\, \hat{g}(\boldsymbol{x})\, dx = \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi})\, e^{i\boldsymbol{x}_0\cdot\boldsymbol{\xi}}\, g(\lambda^{-1}\boldsymbol{\xi})\, d\xi.$$

Taking the limit $\lambda \to \infty$, we obtain

$$f(\boldsymbol{x}_0) \int_{\mathbb{R}^d} \hat{g}(\boldsymbol{x})\, dx = g(\boldsymbol{0}) \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi})\, e^{i\boldsymbol{x}_0\cdot\boldsymbol{\xi}}\, d\xi.$$

Let $g(\boldsymbol{x}) = \exp(-\|\boldsymbol{x}\|^2/2)$ and use the result from Exercise 4.2.3. We then obtain (4.2.11) at an arbitrary point $\boldsymbol{x}_0 \in \mathbb{R}^d$.

We will extend the definition of the Fourier transform to a much broader class of functions, the space of tempered distributions.

Definition 4.2.2 *The space of tempered distributions, $\mathcal{S}'(\mathbb{R}^d)$, is the space of all the continuous linear functionals on $\mathcal{S}(\mathbb{R}^d)$.*

Note that a linear functional T on $\mathcal{S}(\mathbb{R}^d)$ is a tempered distribution if and only if

$$\phi_n \to \phi \text{ in } \mathcal{S}(\mathbb{R}^d) \implies T(\phi_n) \to T(\phi).$$

In the following, we will only consider those tempered distributions that are generated by functions. Then the action of T on ϕ will be written in the form of a duality pairing:

$$T(\phi) = \langle T, \phi \rangle.$$

As an example, any $f \in L^p(\mathbb{R}^d)$, $1 \le p \le \infty$, generates a tempered distribution

$$\mathcal{S}(\mathbb{R}^d) \ni \phi \mapsto \langle f, \phi \rangle = \int_{\mathbb{R}^d} f(\boldsymbol{x})\, \phi(\boldsymbol{x})\, dx. \tag{4.2.12}$$

In this sense, $L^p(\mathbb{R}^d) \subset \mathcal{S}'(\mathbb{R}^d)$.

Recalling the identity (4.2.10), we now define the Fourier transform on $\mathcal{S}'(\mathbb{R}^d)$.

Definition 4.2.3 *Let $f \in \mathcal{S}'(\mathbb{R}^d)$. Then its Fourier transform $\mathcal{F}(f) = \hat{f} \in \mathcal{S}'(\mathbb{R}^d)$ is defined by the formula*

$$\langle \hat{f}, \phi \rangle = \langle f, \hat{\phi} \rangle \quad \forall\, \phi \in \mathcal{S}(\mathbb{R}^d). \tag{4.2.13}$$

It is left as an exercise to show that $\hat{f}$ defined by (4.2.13) belongs to the space $\mathcal{S}'(\mathbb{R}^d)$. Moreover, when $f \in L^1(\mathbb{R}^d)$, the Fourier transform defined by Definition 4.2.3 coincides with the one given in (4.2.3). This can be verified by applying Fubini's theorem.

Notice that Definition 4.2.3 defines the Fourier transform for any $L^p(\mathbb{R}^d)$ function. We mainly use the Fourier transform on $L^2(\mathbb{R}^d)$ related spaces. It can be shown ([73, Chapter 1]) that for $f \in L^2(\mathbb{R}^d)$, its Fourier transform

$$\hat{f}(\boldsymbol{\xi}) = \lim_{R\to\infty} (2\pi)^{-d/2} \int_{\|\boldsymbol{x}\|<R} f(\boldsymbol{x})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, dx \quad \text{in } L^2(\mathbb{R}^d). \tag{4.2.14}$$

Also, we have the inversion formula

$$f(\boldsymbol{x}) = \lim_{R\to\infty} (2\pi)^{-d/2} \int_{\|\boldsymbol{\xi}\|<R} \hat{f}(\boldsymbol{\xi})\, e^{i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\xi \quad \text{in } L^2(\mathbb{R}^d). \tag{4.2.15}$$

We will simply write

$$\hat{f}(\boldsymbol{\xi}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(\boldsymbol{x})\, e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, dx, \tag{4.2.16}$$

$$f(\boldsymbol{x}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi})\, e^{i\boldsymbol{x}\cdot\boldsymbol{\xi}}\, d\xi \tag{4.2.17}$$

even when we only assume $f \in L^2(\mathbb{R}^d)$. Most properties of the Fourier transform on $\mathcal{S}(\mathbb{R}^d)$ carry over to that on $L^2(\mathbb{R}^d)$. For example, we still have the formulas (4.2.7) and (4.2.8), which play key roles in applying the Fourier transform in the study of differential equations. We prove (4.2.7) next, while leaving the proof of (4.2.8) as an exercise.

Assume $f, \partial^\alpha f \in L^2(\mathbb{R}^d)$. Then for any $\phi \in \mathcal{S}(\mathbb{R}^d)$, by Definition 4.2.3,

$$\langle \mathcal{F}(\partial^\alpha f), \phi \rangle = \langle \partial^\alpha f, \mathcal{F}(\phi) \rangle.$$

Performing an integration by part,

$$\langle \mathcal{F}(\partial^\alpha f), \phi \rangle = (-1)^{|\alpha|} \langle f, \partial^\alpha \mathcal{F}(\phi) \rangle.$$

For $\phi \in \mathcal{S}(\mathbb{R}^d)$, we can use (4.2.8). Then,

$$\begin{aligned} \langle \mathcal{F}(\partial^\alpha f), \phi \rangle &= \int_{\mathbb{R}^d} f(\boldsymbol{x})\, \mathcal{F}((i\boldsymbol{\xi})^\alpha \phi(\boldsymbol{\xi}))(\boldsymbol{x})\, dx \\ &= \int_{\mathbb{R}^d} \mathcal{F}(f)(\boldsymbol{\xi})\, (i\boldsymbol{\xi})^\alpha \phi(\boldsymbol{\xi})\, d\xi. \end{aligned}$$

Hence, (4.2.7) holds.

Finally, we quote the following result ([212, Theorem 2.3]).

Theorem 4.2.4 *The Fourier transform operator $\mathcal{F}$ maps $L^2(\mathbb{R}^d)$ onto $L^2(\mathbb{R}^d)$ and is an isometry:*

$$\|\mathcal{F}(f)\|_{L^2(\mathbb{R}^d)} = \|f\|_{L^2(\mathbb{R}^d)} \quad \forall\, f \in L^2(\mathbb{R}^d). \tag{4.2.18}$$

The equality (4.2.18) is called the Plancherel formula, and is the analogue for Fourier transform of Parseval's identity for Fourier series.

Exercise 4.2.1 Prove the properties (4.2.5)–(4.2.8) for functions in $\mathcal{S}(\mathbb{R}^d)$.

Exercise 4.2.2 Verify the property (4.2.9); i.e., for $f \in \mathcal{S}(\mathbb{R}^d)$, show that $\hat{f} \in \mathcal{S}(\mathbb{R}^d)$, and if $f_n \to f$ in $\mathcal{S}(\mathbb{R}^d)$, then $\hat{f}_n \to \hat{f}$ in $\mathcal{S}(\mathbb{R}^d)$.

Exercise 4.2.3 Show that the function $\exp(-\|\boldsymbol{x}\|^2/2) \in \mathcal{S}(\mathbb{R}^d)$; moreover,

$$\mathcal{F}(\exp(-\|\boldsymbol{x}\|^2/2))(\boldsymbol{\xi}) = \exp(-\|\boldsymbol{\xi}\|^2/2),$$

i.e., the Fourier transform operator $\mathcal{F}$ has an eigenvalue 1, with the associated eigenfunction $\exp(-\|\boldsymbol{x}\|^2/2)$.

Exercise 4.2.4 Verify that the mapping (4.2.12) defines a tempered distribution.

Exercise 4.2.5 Show that the formula (4.2.13) defines $\hat{f}$ as a linear and continuous functional on $\mathcal{S}(\mathbb{R}^d)$.

Exercise 4.2.6 Prove (4.2.8) for $f \in L^2(\mathbb{R}^d)$ with $\partial^\alpha \hat{f} \in L^2(\mathbb{R}^d)$.

Exercise 4.2.7 The convolution of two functions on $\mathbb{R}^d$ is defined by the formula

$$(f * g)(\boldsymbol{x}) = \int_{\mathbb{R}^d} f(\boldsymbol{y})\, g(\boldsymbol{x} - \boldsymbol{y})\, dy.$$

Show that if $f, g \in \mathcal{S}(\mathbb{R}^d)$, then $f * g \in \mathcal{S}(\mathbb{R}^d)$ and

$$\mathcal{F}(f * g) = \mathcal{F}(f)\, \mathcal{F}(g).$$

Exercise 4.2.8 Prove (4.2.18) for $f \in \mathcal{S}(\mathbb{R}^d)$.
Hint: Introduce $g(\boldsymbol{x}) = \overline{f(-\boldsymbol{x})}$ and $h = f * g$. Then apply the result from Exercise 4.2.7 and note

$$h(\mathbf{0}) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathcal{F}(h)(\boldsymbol{y})\, dy.$$

Exercise 4.2.9 Show that (4.2.18) is equivalent to the identity

$$\int_{\mathbb{R}^d} f(\boldsymbol{x})\, \overline{g(\boldsymbol{x})}\, dx = \int_{\mathbb{R}^d} \hat{f}(\boldsymbol{\xi})\, \overline{\hat{g}(\boldsymbol{\xi})}\, d\xi \quad \forall\, f, g \in L^2(\mathbb{R}^d).$$

Hint: Derive and use the identity

$$\begin{aligned}\int_{\mathbb{R}^d} f(\boldsymbol{x})\, \overline{g(\boldsymbol{x})}\, dx = \frac{1}{4}\left[\|f+g\|^2_{L^2(\mathbb{R}^d)} - \|f-g\|^2_{L^2(\mathbb{R}^d)}\right] \\ + \frac{1}{4i}\left[\|f-ig\|^2_{L^2(\mathbb{R}^d)} - \|f+ig\|^2_{L^2(\mathbb{R}^d)}\right]\end{aligned}$$

for any $f, g \in L^2(\mathbb{R}^d)$.

4.3 Discrete Fourier transform

The Fourier series and Fourier transform are important mathematical tools in the analysis of many problems. However, they are not suitable for computer implementation due to the presence of integrals in the formulas. That is why we also introduce the discrete Fourier transform.

Recall the complex form of the Fourier series (4.1.4) with the coefficients given by (4.1.5). Since f is a 2π-periodic function, we can express the Fourier coefficients as

$$c_k = \frac{1}{2\pi}\int_0^{2\pi} f(x)\, e^{-ikx} dx, \quad k \text{ integer}.$$

Let us apply the trapezoidal rule to approximate the integration. For this purpose, let n be a natural number, and decompose the interval $[0, 2\pi]$ into n equal sub-intervals with the nodes $x_j = jh$, $0 \le j \le n$, $h = 2\pi/n$. Recalling that $f(0) = f(2\pi)$, we have the approximate formula

$$c_k \approx \frac{1}{n}\sum_{j=0}^{n-1} f(x_j)\, e^{-ikx_j}.$$

Since n function values, $\{f(x_j)\}_{j=0}^{n-1}$, are used in computing (approximately) the Fourier coefficients, it is natural to try to use n coefficients to recover the n function values.

Definition 4.3.1 *Let n be a positive integer, and let $\{y_j\}_{j=0}^{n-1} \subset \mathbb{C}$ be a sequence of complex numbers. Then the discrete Fourier transform is the sequence $\{\hat{y}_k\}_{k=0}^{n-1} \subset \mathbb{C}$ defined by the formula*

$$\hat{y}_k = \sum_{j=0}^{n-1} \omega_n^{-kj} y_j, \quad 0 \le k \le n-1, \tag{4.3.1}$$

where

$$\omega_n = e^{2\pi i/n}. \tag{4.3.2}$$

We call n the order of the discrete Fourier transform.

To express the discrete Fourier transform with the matrix/vector notation, we introduce the vectors

$$\boldsymbol{y} = (y_0, \ldots, y_{n-1})^T, \quad \hat{\boldsymbol{y}} = (\hat{y}_0, \ldots, \hat{y}_{n-1})^T$$

and the matrix

$$F_n = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^4 & \cdots & \omega_n^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_n^{n-1} & \omega_n^{2(n-1)} & \cdots & \omega_n^{(n-1)^2} \end{pmatrix}.$$

Then the discrete Fourier transform is

$$\hat{\boldsymbol{y}} = \overline{F_n}\boldsymbol{y},$$

where $\overline{F_n}$ is the matrix obtained from F_n by taking the conjugate of its elements.

We can determine the original sequence $\{y_j\}_{j=0}^{n-1}$ from $\{\hat{y}_j\}_{j=0}^{n-1}$.

Theorem 4.3.2 *The inverse of the discrete Fourier transform matrix is*

$$(\overline{F_n})^{-1} = \frac{1}{n} F_n. \tag{4.3.3}$$

Consequently, the inverse discrete Fourier transform $\boldsymbol{y} = (\overline{F_n})^{-1}\hat{\boldsymbol{y}}$ *is*

$$y_j = \frac{1}{n}\sum_{k=0}^{n-1} \omega_n^{jk}\hat{y}_k, \quad 0 \le j \le n-1. \tag{4.3.4}$$

Proof. We only need to verify

$$\overline{F_n}F_n = n\, I,$$

which follows from the following summation formula: for $j, l = 0, \cdots, n-1$,

$$\sum_{k=0}^{n-1} \omega_n^{kl}\omega_n^{-kj} = \begin{cases} n, & \text{if } j = l, \\ 0, & \text{otherwise.} \end{cases} \tag{4.3.5}$$

Proof of this formula is elementary and is left as Exercise 4.3.1. □

The fast Fourier transform (FFT) (see [59]) is a fast way to compute the multiplication of a vector by the discrete Fourier transform matrix. There were earlier versions of the discrete Fourier transform and the FFT, dating back to, at least, Gauss. We first explain the idea of the FFT by exploring relations between discrete Fourier transforms of different orders. For convenience, we will also use the notation $\mathcal{F}_n(\{y_j\}_{j=0}^{n-1})$ to represent the vector obtained from the vector $(y_0, \ldots, y_{n-1})^T$ through the discrete Fourier transform. Let n be a positive integer, and let $\boldsymbol{y} = (y_0, \ldots, y_{2n-1})^T$ be a $(2n)$-dimensional vector. The purpose is to figure out how to compute

$$\hat{\boldsymbol{y}} = \mathcal{F}_{2n}(\{y_j\}_{j=0}^{2n-1}) \tag{4.3.6}$$

efficiently. The components of $\hat{\boldsymbol{y}}$ are given by

$$\hat{y}_k = \sum_{j=0}^{2n-1} \omega_{2n}^{-kj} y_j, \quad 0 \le k \le 2n-1. \tag{4.3.7}$$

Since

$$\omega_n = e^{2\pi i/n} = \omega_{2n}^2,$$

we can rewrite $\hat{y}_k$ as

$$\begin{aligned}\hat{y}_k &= \sum_{j=0}^{n-1} \omega_n^{-kj} y_{2j} + \omega_{2n}^{-k} \sum_{j=0}^{n-1} \omega_n^{-kj} y_{2j+1} \\ &= \mathcal{F}_n(\{y_{2j}\}_{j=0}^{n-1})_k + \omega_{2n}^{-k} \mathcal{F}_n(\{y_{2j+1}\}_{j=0}^{n-1})_k\end{aligned}$$

for $k = 0, \ldots, 2n-1$. Cost saving comes from the observation that

$$\omega_n^{-(k+n)j} = \omega_n^{-kj}, \quad \omega_{2n}^{-(k+n)} = -\omega_{2n}^{-k}$$

and so

$$\hat{y}_k = \mathcal{F}_n(\{y_{2j}\}_{j=0}^{n-1})_k + \omega_{2n}^{-k} \mathcal{F}_n(\{y_{2j+1}\}_{j=0}^{n-1})_k, \quad 0 \le k \le n-1, \quad (4.3.8)$$

$$\hat{y}_{k+n} = \mathcal{F}_n(\{y_{2j}\}_{j=0}^{n-1})_k - \omega_{2n}^{-k} \mathcal{F}_n(\{y_{2j+1}\}_{j=0}^{n-1})_k, \quad 0 \le k \le n-1. \quad (4.3.9)$$

Notice that for $k = 0, \cdots, n-1$, the calculation of the components $\hat{y}_k$ and $\hat{y}_{k+n}$ share the common terms $\mathcal{F}_n(\{y_{2j}\}_{j=0}^{n-1})_k$ and $\mathcal{F}_n(\{y_{2j+1}\}_{j=0}^{n-1})_k$. Let us count the numbers of multiplications needed in computing the discrete Fourier transform vector $\hat{\boldsymbol{y}}$ directly from (4.3.7) and from (4.3.8)–(4.3.9). Assuming factors of the forms ω_{2n}^{-kj} and ω_n^{-kj} have been stored for use. Then the formula (4.3.7) requires $(2n)^2 = 4n^2$ multiplications. Suppose the order n transform vectors $\mathcal{F}_n(\{y_{2j}\}_{j=0}^{n-1})$ and $\mathcal{F}_n(\{y_{2j+1}\}_{j=0}^{n-1})$ are computed directly from the definition. Then each requires n^2 multiplications. In (4.3.8) and (4.3.9), there is one additional multiplication by ω_{2n}^{-k}. Therefore, the computation based on (4.3.8)–(4.3.9) requires $2n^2 + n$ multiplications, or roughly half of the multiplications required by (4.3.7), when n is large. A similar argument shows that the number of additions required is also almost halved with the use of (4.3.8)–(4.3.9).

The above procedure can be applied repeated as long as the order of the discrete Fourier transform is an even number. The resulting algorithm is called the *fast Fourier transform* (FFT). Suppose the order of the transform is $n = 2^m$. We can then apply the above procedure m times, and the last iteration in the procedure involves some order one discrete Fourier transforms, i.e., some components of the given vector. Let N_m denote the number of multiplications required for an order $n(= 2^m)$ FFT. Then

$$N_m = 2\, N_{m-1} + 2^{m-1}.$$

From this,

$$N_m = 2^2 N_{m-2} + 2 \cdot 2^{m-1}.$$

By an induction, for $j = 1, \ldots, m$,

$$N_m = 2^j N_{m-j} + j \cdot 2^{m-1}.$$

Since no multiplication is needed for order one discrete Fourier transforms, $N_0 = 0$, and we have

$$N_m = 2^m N_0 + m\, 2^{m-1} = m\, 2^{m-1}.$$

So the total number of multiplications for FFT is

$$N_m = m\, 2^{m-1} = 0.5\, n\, \log_2 n.$$

This is compared to n^2 multiplications if the order n discrete Fourier transform is computed directly.

We now briefly comment on the discrete Fourier transform in higher dimensions. The two-dimensional discrete Fourier transform is defined as follows. Let n_1 and n_2 be two positive integers, and let

$$\{y_{j_1 j_2} \mid 0 \le j_1 \le n_1 - 1,\, 0 \le j_2 \le n_2 - 1\} \subset \mathbb{C}$$

be a two dimensional array of complex numbers. Then the discrete Fourier transform is the two-dimensional array

$$\{\hat{y}_{k_1 k_2} \mid 0 \le k_1 \le n_1 - 1,\, 0 \le k_2 \le n_2 - 1\} \subset \mathbb{C}$$

defined by the formula

$$\hat{y}_{k_1 k_2} = \sum_{j_1=0}^{n_1-1} \sum_{j_2=0}^{n_2-1} \omega_{n_1}^{-k_1 j_1} \omega_{n_2}^{-k_2 j_2} y_{j_1 j_2}, \quad 0 \le k_1 \le n_1 - 1,\, 0 \le k_2 \le n_2 - 1. \tag{4.3.10}$$

The inverse discrete Fourier transform formula is (its verification is left as Exercise 4.3.5)

$$y_{j_1 j_2} = \frac{1}{n_1 n_2} \sum_{k_1=0}^{n_1-1} \sum_{k_2=0}^{n_2-1} \omega_{n_1}^{j_1 k_1} \omega_{n_2}^{j_2 k_2} \hat{y}_{k_1 k_2}, \quad 0 \le j_1 \le n_1 - 1,\, 0 \le j_2 \le n_2 - 1. \tag{4.3.11}$$

We notice that the formula (4.3.10) can be rewritten as

$$\hat{y}_{k_1 k_2} = \sum_{j_1=0}^{n_1-1} \omega_{n_1}^{-k_1 j_1} \left(\sum_{j_2=0}^{n_2-1} \omega_{n_2}^{-k_2 j_2} y_{j_1 j_2} \right), \quad 0 \le k_1 \le n_1 - 1,\, 0 \le k_2 \le n_2 - 1,$$

i.e., the two-dimensional discrete Fourier transform can be computed by iterated one-dimensional discrete Fourier transforms. The same observation applies to the two-dimensional inverse discrete Fourier transform. These are true for higher dimensional direct and inverse discrete Fourier transforms. For this reason, it is usually sufficient to focus the analysis on the one-dimensional case only.

Finally, we remark that unlike for the continuous case, the discrete Fourier transform and its inverse always exist for any given sequence of numbers.

Exercise 4.3.1 Prove the formula (4.3.5) for $j, l = 0, \cdots, n-1$.

Exercise 4.3.2 For the discrete Fourier transform defined by (4.3.1), prove the discrete Parseval's equality:

$$\sum_{k=0}^{n-1} |\hat{y}_k|^2 = n \sum_{j=0}^{n-1} |y_j|^2.$$

Exercise 4.3.3 Show that the matrix form of (4.3.8)–(4.3.9) is

$$\overline{F_{2n}}\boldsymbol{y} = \begin{pmatrix} I_n & \overline{D_n} \\ I_n & -\overline{D_n} \end{pmatrix} \begin{pmatrix} \overline{F_n} & 0 \\ 0 & \overline{F_n} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_e \\ \boldsymbol{y}_o \end{pmatrix},$$

where I_n is the identity matrix of order n, $D_n = \mathrm{diag}(1, \omega_{2n}, \cdots, \omega_{2n}^{n-1})$ is a diagonal matrix of order n, $\boldsymbol{y} = (y_0, y_1, \cdots, y_{2n-1})^T$, $\boldsymbol{y}_e = (y_0, y_2, \cdots, y_{2n-2})^T$, and $\boldsymbol{y}_o = (y_1, y_3, \cdots, y_{2n-1})^T$.

Exercise 4.3.4 Show that for the inverse discrete Fourier transform, the analog of (4.3.8)–(4.3.9) is

$$y_j = \frac{1}{2}\left[\mathcal{F}_n^{-1}(\{\hat{y}_{2k}\}_{k=0}^{n-1})_j + \omega_{2n}^j \mathcal{F}_n^{-1}(\{\hat{y}_{2k+1}\}_{k=0}^{n-1})_j\right], \quad 0 \le j \le n-1,$$
$$y_{j+n} = \frac{1}{2}\left[\mathcal{F}_n^{-1}(\{\hat{y}_{2k}\}_{k=0}^{n-1})_j - \omega_{2n}^j \mathcal{F}_n^{-1}(\{\hat{y}_{2k+1}\}_{k=0}^{n-1})_j\right], \quad 0 \le j \le n-1.$$

Exercise 4.3.5 Derive the inverse discrete Fourier transform formula (4.3.11).

4.4 Haar wavelets

The Fourier series is most suitable for approximations of periodic smooth functions. To approximate functions of general features (non-periodic, non-smooth either locally or globally), especially in such application areas as signal processing, wavelets provide a better tool. The word wavelet means small wave. Wavelets are an alternative to the Fourier transform to represent a signal, using short wavelets instead of long waves. They are also the basis of many image compression algorithms.

In this section and the next, our discussion of the subject is for real valued functions only. The extension to the complex case is straightforward.

From a functional analysis perspective, the goal is to introduce a decomposition of a function space into subspaces so that each function in the space can be decomposed into pieces in the subspaces. The basic idea of a wavelet analysis is to generate building blocks for the space decomposition through translations and dilations of a single function called a scaling function. In a wavelet analysis, a function is hierarchically decomposed, and coefficients corresponding to a certain level reflect details of the function at that level. In this section, we consider the simplest wavelets, the

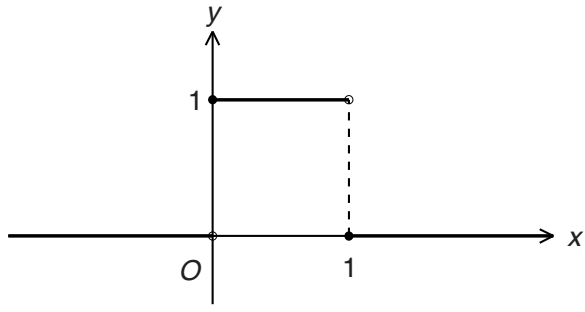

FIGURE 4.5. Haar scaling function

Haar wavelets. The corresponding scaling function is the basic unit step function:

$$\phi(x) = \begin{cases} 1, & 0 \le x < 1, \\ 0, & x < 0 \text{ or } x \ge 1. \end{cases} \tag{4.4.1}$$

Figure 4.5 shows the graph of the Haar scaling function.

Integer translations of $\phi(x)$ are functions of the form $\phi(x-k)$, $k \in \mathbb{Z}$. Recall that $\mathbb{Z}$ is the set of all the integers. Let V_0 denote the space of all finite linear combinations of $\{\phi(x-k) \mid k \in \mathbb{Z}\}$. In other words, $f \in V_0$ if and only if f has a bounded support on $\mathbb{R}$ and is a constant on any interval of the form $[k, k+1)$, $k \in \mathbb{Z}$. A general expression for $f \in V_0$ is

$$f(x) = \sum_k^{\circ} a_k \phi(x-k),$$

where $\sum_k^{\circ}$ stands for a summation for a finite number of k's. We comment that the use of a finite summation in defining the functions in V_0 is for convenience in computation and is not an essential restriction. It is also possible to define V_0 to be the space of all the functions of the form

$$f(x) = \sum_k a_k \phi(x-k),$$

where $\sum_k |a_k|^2 < \infty$ so that $f \in L^2(\mathbb{R})$.

Now for any integer j (positive or negative), we use all finite linear combinations of scaled functions $\phi(2^j x - k)$, $k \in \mathbb{Z}$, to form the level j subspace V_j of $L^2(\mathbb{R})$. A general function in V_j has the form

$$f(x) = \sum_k^{\circ} a_k \phi(2^j x - k).$$

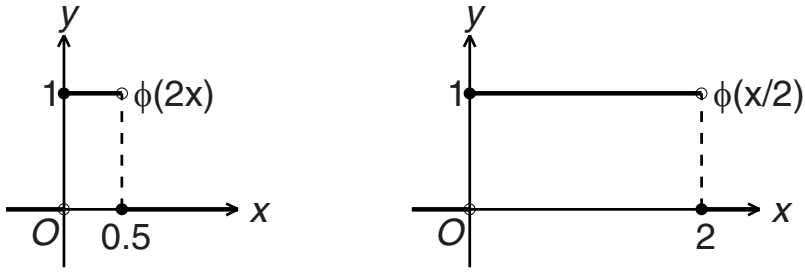

FIGURE 4.6. Scaled Haar functions $\phi(2x)$ and $\phi(x/2)$

Note that $\phi(2^j x - k)$ equals 1 for $x \in [k\,2^{-j}, (k+1)\,2^{-j})$, and 0 for x elsewhere. Figure 4.6 shows the graphs of the scaled functions $\phi(2x)$ and $\phi(2^{-1}x)$.

A number is called a dyadic number if it is an integer multiple of an integer power of 2. Denote the set of all diadic numbers by $\mathbb{D}$. For each integer j (positive, negative, or zero), denote the set of all integer multiples of 2^{-j} by $\mathbb{D}_j$. An interval of the form $[k\,2^{-j}, (k+1)\,2^{-j})$ is called a dyadic interval of level j. We see that $f \in V_j$ is a piecewise constant function and is constant on any dyadic interval of level j. The subspaces V_j, $j \in \mathbb{Z}$, are called Haar scaling spaces. Their basic properties are given in the next theorem.

Theorem 4.4.1 *For the subspaces V_j, $j \in \mathbb{Z}$, defined through the translations and scaling of the Haar scaling function, the following statements hold.*

(1) *For any $j \in \mathbb{Z}$, $\{2^{j/2}\phi(2^j x - k) \mid k \in \mathbb{Z}\}$ is an orthonormal basis of V_j.*

(2) *$f(x) \in V_j$ if and only if $f(2^{-j}x) \in V_0$.*

(3) *$V_j \subsetneqq V_{j+1}$.*

(4) *$\overline{\cup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R})$, where the closure is taken with respect to the $L^2(\mathbb{R})$-norm.*

(5) *$\cap_{j\in\mathbb{Z}} V_j = \{0\}$.*

The first three properties are straightforward to verify. Property (1) is a *shift-invariance property*; it requires the scaling function $\phi(x) \in L^2(\mathbb{R})$ and its translates $\phi(x-k)$, $k \neq 0$ integer, to be linearly independent and form an orthonormal basis for V_0. Property (2) is a *scale-invariance property*; combined with the definition of V_0, this property can be used to define any other subspaces V_j, $j \in \mathbb{Z}$. Property (3) says the sequence of the subspaces is nested:

$$\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots .$$

Property (4) can be shown through the density of continuous functions in $L^2(\mathbb{R})$ and the approximability of a continuous function by piecewise constants in $L^2(\mathbb{R})$. Property (5) is called a *separation condition*. For its proof, note that if $f \in \cap_{j\in\mathbb{Z}} V_j$, then $f \in V_{-j}$ and so $f(x)$ is constant on $[-2^j, 0)$ and $[0, 2^j)$. Letting $j \to \infty$, the requirement $f \in L^2(\mathbb{R})$ implies $f(x) \equiv 0$. Details of the proof are left as Exercise 4.4.2.

With the nested sequence of subspaces

$$V_0 \subset V_1 \subset V_2 \subset \cdots \subset L^2(\mathbb{R}),$$

a *multiresolution approach* to approximating a function f is as follows. First, define an approximation $f_0 \in V_0$ of f from the lowest level subspace V_0. Such an approximation can approximate f well only in regions where the graph of f is rather flat in intervals of the form $[k, k+1)$, $k \in \mathbb{Z}$. To improve the approximation quality, we supplement f_0 by a $w_0 \in V_1$ so that $f_1 = f_0 + w_0 \in V_1$ approximates f well in half-size intervals of the form $[k/2, (k+1)/2)$, $k \in \mathbb{Z}$, where f changes slowly. For the procedure to be efficient, we choose w_0 to be orthogonal to V_0. This is an essential step. The process is then repeated to generate higher level approximations

$$f_l = f_0 + \sum_{j=0}^{l-1} w_j,$$

where w_j is orthogonal to V_j.

Thus, for any $j \in \mathbb{Z}$, we consider the orthogonal decomposition

$$V_{j+1} = V_j \oplus W_j$$

and want to identify the orthogonal complement W_j.

Theorem 4.4.2 *A function*

$$f(x) = \sum_k^{\smile} a_k \phi(2^{j+1}x - k) \in V_{j+1}$$

is orthogonal to V_j if and only if

$$a_{2k+1} = -a_{2k} \quad \forall\, k \in \mathbb{Z}.$$

Thus, such a function has the form

$$f(x) = \sum_k^{\smile} a_{2k} \left[\phi(2^{j+1}x - 2k) - \phi(2^{j+1}x - 2k - 1)\right].$$

Proof. The function f is orthogonal to V_j if and only if f is orthogonal to $\phi(2^j x - k)$ for any $k \in \mathbb{Z}$. Note that $\phi(2^j x - k)$ equals 1 for $x \in [k\,2^{-j}, (k+$

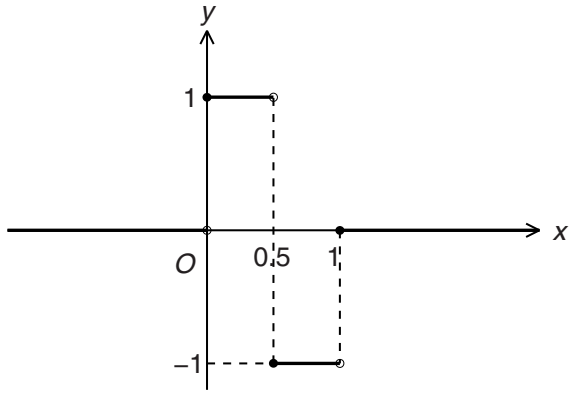

FIGURE 4.7. Haar wavelet function

$1)\,2^{-j})$ and is 0 otherwise. We have

$$\begin{aligned}\int_{\mathbb{R}} f(x)\,\phi(2^j x-k)\,dx &= \int_{k\,2^{-j}}^{(k+1)\,2^{-j}} f(x)\,dx \\ &= \int_{2k\,2^{-(j+1)}}^{(2k+1)\,2^{-(j+1)}} a_{2k}\,dx + \int_{(2k+1)\,2^{-(j+1)}}^{(2k+2)\,2^{-(j+1)}} a_{2k+1}\,dx \\ &= (a_{2k}+a_{2k+1})\,2^{-(j+1)}.\end{aligned}$$

Therefore, f is orthogonal to V_j if and only if $a_{2k+1}+a_{2k}=0$ for any $k\in\mathbb{Z}$. □

Theorem 4.4.2 suggests the definition of the function

$$\psi(x)=\phi(2x)-\phi(2x-1), \tag{4.4.2}$$

called the Haar wavelet function. Its graph is shown in Figure 4.7. As a consequence of Theorem 4.4.2, we conclude that the orthogonal complement $W_j = V_{j+1}\ominus V_j$ consists of all finite linear combinations of $\psi(2^j x-k)$, $k\in\mathbb{Z}$. The subspaces W_j, $j\in\mathbb{Z}$, are called the Haar wavelet spaces.

For any two integers $j>i$, we have the following sequence of orthogonal decompositions:

$$\begin{aligned}V_j &= V_{j-1}\oplus W_{j-1}\\ &= V_{j-2}\oplus W_{j-2}\oplus W_{j-1}\\ &= \cdots\\ &= V_i\oplus W_i\oplus\cdots\oplus W_{j-1}.\end{aligned}$$

The properties (4) and (3) of Theorem 4.4.1 imply that for any $j\in\mathbb{Z}$,

$$L^2(\mathbb{R}) = V_j\oplus W_j\oplus\cdots.$$

In particular, with $j = 0$, we have

$$L^2(\mathbb{R}) = V_0 \oplus W_0 \oplus W_1 \oplus \cdots,$$

and any $f \in L^2(\mathbb{R})$ has the unique expansion

$$f = f_0 + \sum_{j=0}^{\infty} w_j = f_0 + \lim_{n\to\infty} \sum_{j=0}^{n} w_j \quad \text{in } L^2(\mathbb{R})$$

for some $f_0 \in V_0$ and $w_j \in W_j$, $j \geq 0$.

Given a function $f(x)$, which represents a signal for example, we can approximate/process it via the Haar wavelet system by the following procedure. First, we choose a positive integer J so that the main features of f over intervals of width 2^{-J} are to be captured. Then we compute the function values

$$a_k^J = f(k\,2^{-J}).$$

The integer variable k ranges over a set for which the points of the form $(k\,2^{-J})$ belong to the interval where $f(x)$ is to be approximated. We define the highest-level approximation

$$f_J(x) = \breve{\sum_k} a_k^J \phi(2^J x - k), \tag{4.4.3}$$

where the summation is over the range of k.

Next, we rewrite $f_J(x)$ in the form

$$f_J(x) = f_0(x) + \sum_{j=0}^{J-1} w_j(x), \quad w_j \in W_j,\ 0 \leq j \leq J-1. \tag{4.4.4}$$

This decomposition is unique. We emphasize that the same discussion applies if we decide the decomposition starts from a different level, say level l:

$$f_J(x) = f_l(x) + \sum_{j=l}^{J-1} w_j(x), \quad w_j \in W_j,\ l \leq j \leq J-1.$$

For definiteness, in the following, we use (4.4.4), i.e., the lowest-level approximation takes place in the level zero subspace V_0. The decomposition (4.4.4) is constructed based on the following formulas: for $j, k \in \mathbb{Z}$,

$$\phi(2^j x - 2k) = \frac{1}{2}\left[\phi(2^{j-1}x - k) + \psi(2^{j-1}x - k)\right], \tag{4.4.5}$$

$$\phi(2^j x - 2k - 1) = \frac{1}{2}\left[\phi(2^{j-1}x - k) - \psi(2^{j-1}x - k)\right]. \tag{4.4.6}$$

These formulas can be derived from the relations

$$\phi(2x) = \frac{1}{2}\left[\phi(x) + \psi(x)\right],$$
$$\phi(2x-1) = \frac{1}{2}\left[\phi(x) - \psi(x)\right].$$

Proof of (4.4.5)–(4.4.6) is left as Exercise 4.4.3.

We have the following Haar decomposition theorem.

Theorem 4.4.3 *The function*

$$f_j(x) = \breve{\sum_k} a_k^j \phi(2^j x - k)$$

can be decomposed as

$$f_j(x) = f_{j-1}(x) + w_{j-1}(x),$$

where

$$f_{j-1}(x) = \breve{\sum_k} a_k^{j-1} \phi(2^{j-1}x - k) \in V_{j-1},$$
$$w_{j-1}(x) = \breve{\sum_k} b_k^{j-1} \psi(2^{j-1}x - k) \in W_{j-1}$$

with the coefficients

$$a_k^{j-1} = \frac{1}{2}\left(a_{2k}^j + a_{2k+1}^j\right), \tag{4.4.7}$$
$$b_k^{j-1} = \frac{1}{2}\left(a_{2k}^j - a_{2k+1}^j\right). \tag{4.4.8}$$

Proof. We write

$$f_j(x) = \breve{\sum_k}\left[a_{2k}^j \phi(2^j x - 2k) + a_{2k+1}^j \phi(2^j x - 2k - 1)\right].$$

Using the formulas (4.4.5)–(4.4.6), we obtain

$$f_j(x) = \breve{\sum_k}\left\{a_{2k}^j \frac{1}{2}\left[\phi(2^{j-1}x - k) + \psi(2^{j-1}x - k)\right] \right.$$
$$\left. + a_{2k+1}^j \frac{1}{2}\left[\phi(2^{j-1}x - k) - \psi(2^{j-1}x - k)\right]\right\}.$$

Rearrange the terms to get the specified decomposition. □

Applying Theorem 4.4.3 with $j = J, \cdots, 1$, we obtain the decomposition (4.4.4) with

$$f_0(x) = \breve{\sum_k} a_k^0 \phi(x-k), \tag{4.4.9}$$

$$w_j(x) = \breve{\sum_k} b_k^j \psi(2^j x - k), \quad 0 \le j \le J-1, \tag{4.4.10}$$

where the coefficients $b_k^{J-1}, \cdots, b_k^0, a_k^0$ are computed recursively: for $j = J, \cdots, 1$, we apply the formulas (4.4.7) and (4.4.8).

Now that the decomposition (4.4.4), together with (4.4.9)–(4.4.10), is available, some of the coefficients can be modified depending on the needs. Suppose $f(x)$ represents a signal. Then for large values of j, $w_j(x)$ contains high frequency components, and can be viewed as noise. If the purpose of processing the signal is to remove the noise, then the coefficients corresponding to those high frequencies are set to zero. If the purpose is to compress the data, then those coefficients with their absolute values below certain threshold are set to zero, and doing this does not severely influence the accuracy of approximation. So after the processing, we get a new approximation from (4.4.3):

$$\tilde{f}_J(x) = \tilde{f}_0(x) + \sum_{j=0}^{J-1} \tilde{w}_j(x), \tag{4.4.11}$$

where

$$\tilde{f}_0(x) = \breve{\sum_k} \tilde{a}_k^0 \phi(x-k), \tag{4.4.12}$$

$$\tilde{w}_j(x) = \breve{\sum_k} \tilde{b}_k^j \psi(2^j x - k), \quad 0 \le j \le J-1. \tag{4.4.13}$$

The function $\tilde{f}_J$ will be used as the approximation of the given function $f(x)$. The computation of $\tilde{f}_J(x)$ is more efficiently done when it is written in terms of the basis functions $\phi(2^J x - k)$, $k \in \mathbb{Z}$. This step is called reconstruction. From (4.4.5)–(4.4.6), we obtain the following formulas:

$$\phi(2^j x - k) = \phi(2^{j+1}x - 2k) + \phi(2^{j+1}x - 2k - 1), \tag{4.4.14}$$

$$\psi(2^j x - k) = \phi(2^{j+1}x - 2k) - \phi(2^{j+1}x - 2k - 1). \tag{4.4.15}$$

Using these formulas, we can easily show the following Haar reconstruction theorem.

Theorem 4.4.4 *The function*

$$\tilde{f}_{j+1}(x) = \breve{\sum_k} \tilde{a}_k^j \phi(2^j x - k) + \breve{\sum_k} \tilde{b}_k^j \psi(2^j x - k)$$

can be expressed as

$$\tilde{f}_{j+1}(x) = \breve{\sum_k} \tilde{a}_k^{j+1} \phi(2^{j+1}x - k)$$

where

$$\tilde{a}_{2k}^{j+1} = \tilde{a}_k^j + \tilde{b}_k^j, \tag{4.4.16}$$

$$\tilde{a}_{2k+1}^{j+1} = \tilde{a}_k^j - \tilde{b}_k^j. \tag{4.4.17}$$

Applying Theorem 4.4.4 with $j = 0, 1, \cdots, J-1$, we finally get the expression

$$\tilde{f}_J(x) = \breve{\sum_k} \tilde{a}_k^J \phi(2^J x - k) \tag{4.4.18}$$

where the coefficients $\tilde{a}_k^J$ are computed recursively from (4.4.16)–(4.4.17). Formula (4.4.18) is convenient to use in computing the approximate function values.

Exercise 4.4.1 Given $f \in L^2(0,1)$, determine the coefficients a_0 and a_1 in the function

$$g(x) = a_0\phi(2x) + a_1\phi(2x-1)$$

so that $\|f-g\|_{L^2(0,1)}$ is minimal. Note that we get the same values for a_0 and a_1 when we minimize $\|f-g\|_{L^2(0,1/2)}$ and $\|f-g\|_{L^2(1/2,1)}$ separately.

Exercise 4.4.2 Prove Theorem 4.4.1.

Exercise 4.4.3 Derive the formulas (4.4.5) and (4.4.6).

Exercise 4.4.4 Show that the rescaled Haar wavelets $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$ form an orthonormal basis of $L^2(\mathbb{R})$:

$$\int_{\mathbb{R}} \psi_{jk}(x)\, \psi_{lm}(x)\, dx = \delta_{jl}\delta_{km}.$$

4.5 Multiresolution analysis

The Haar wavelets discussed in Section 4.4 are the simplest wavelets. A general framework for the theory of wavelets, called multiresolution analysis, was developed in [163]. For general wavelets, the discussion starts with the scaling function ϕ and scaling spaces V_j, $j \in \mathbb{Z}$. Recall Theorem 4.4.1 for the basic properties of the subspaces associated with the Haar wavelets.

Definition 4.5.1 *A sequence of subspaces of $L^2(\mathbb{R})$, $\{V_j \mid j \in \mathbb{Z}\}$, is called a multiresolution analysis with scaling function ϕ if the following conditions holds:*

(1) (Shift-invariance) $\{\phi(x-k) \mid k \in \mathbb{Z}\}$ *is an orthonormal basis of* V_0.

(2) (Scale-invariance) $f(x) \in V_j$ *if and only if* $f(2^{-j}x) \in V_0$.

(3) (Nested sequence) $V_j \subset V_{j+1}$.

(4) (Density) $\overline{\cup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R})$, *where the closure is taken with respect to the* $L^2(\mathbb{R})$*-norm.*

(5) (Separation) $\cap_{j\in\mathbb{Z}} V_j = \{0\}$.

The orthogonality condition in (1) may be weakened to the requirement that $\{\phi(x-k) \mid k \in \mathbb{Z}\}$ forms a stable basis: Any function $f \in V_0$ can be written uniquely as

$$f(x) = \sum_{k\in\mathbb{Z}} f_k \phi(x-k) \quad \text{in } L^2(\mathbb{R})$$

and

$$c_0 \sum_{k\in\mathbb{Z}} |f_k|^2 \le \|f\|^2_{L^2(\mathbb{R})} \le c_1 \sum_{k\in\mathbb{Z}} |f_k|^2$$

for some constants $0 < c_0 \le c_1 < \infty$ independent of $f \in V_0$. Since the orthogonality usually helps simplify calculations dramatically, it is popular to require the orthonormality condition for the scaling function and its translates. Moreover, when the orthonormality for ϕ is not valid, it is possible to define a new scaling function from ϕ such that the orthonormality condition holds ([234]). Condition (4) requires the density of the subspace $\cup_{j=-\infty}^{\infty} V_j$ in $L^2(\mathbb{R})$: for any $f \in L^2(\mathbb{R})$, there exists a sequence $\{f_n\} \subset \cup_{j=-\infty}^{\infty} V_j$ such that

$$\|f - f_n\|_{L^2(\mathbb{R})} \to 0 \quad \text{as } n \to \infty.$$

Combined with the condition (3), we can restate this as follows: for any $f \in L^2(\mathbb{R})$, there exist $f_n \in V_n$, $n = 0, 1, 2, \cdots$, such that

$$\|f - f_n\|_{L^2(\mathbb{R})} \to 0 \quad \text{as } n \to \infty.$$

The next result follows immediately from Definition 4.5.1, and its proof is left as an exercise.

Proposition 4.5.2 *For a multiresolution analysis with scaling function* ϕ, $\{2^{j/2}\phi(2^j x - k) \mid k \in \mathbb{Z}\}$ *is an orthonormal basis of* V_j.

Wavelet spaces W_j are constructed as the difference between V_{j+1} and V_j such that W_j is orthogonal to V_j. Since for $k < j$, W_k is a subspace of V_j, W_k and W_j are orthogonal. So the wavelet spaces are mutually orthogonal. The sequence of the wavelet spaces $\{W_n\}$ has properties similar to the conditions required in the definition of the multiresolution analysis. Specifically, we mention the following.

(1) There is a function $\psi \in V_1$ such that $\{\psi(x-k) \mid k \in \mathbb{Z}\}$ is an orthonormal basis of W_0.

(2) $f(x) \in W_j$ if and only if $f(2^{-j}x) \in W_0$.

(3) $W_j \perp W_k$ for $j \neq k$.

(4) $\overline{\cup_{j\in\mathbb{Z}} W_j} = L^2(\mathbb{R})$, where the closure is taken with respect to the $L^2(\mathbb{R})$-norm.

From Theorem 4.4.1, we see that the Haar system is a multiresolution analysis. Moreover, in the Haar system, any subspace V_j consists of functions with bounded support.

The Haar wavelet has a compact support. The major disadvantage of the Haar functions is the lack of smoothness: a general function in the Haar subspaces is discontinuous. A consequence of this is the poor approximation quality for continuous functions. A main task is then to construct a multiresolution analysis consisting of smoother functions. In [62], Daubechies showed the possibility of constructing other wavelets with compact support. For the Haar wavelets, we have the simple explicit formulas, and properties of the Haar wavelets are verified directly. For other wavelets, generally there are no elementary formulas for the scaling function ϕ and wavelet function ψ; instead, the functions are determined implicitly by a *dilation equation* and a *wavelet equation*, respectively. Properties of the wavelets, such as the support interval, orthogonality, and smoothness, are then determined from these equations.

The starting point in the construction of the scaling function ϕ is the scaling equation

$$\phi(x) = \sum_k p_k \sqrt{2}\, \phi(2x-k) \tag{4.5.1}$$

for some coefficients p_k, $k \in \mathbb{Z}$. This relation holds because $\phi \in V_0 \subset V_1$ and $\{\sqrt{2}\phi(2x-k) \mid k \in \mathbb{Z}\}$ is an orthonormal basis. Note that if ϕ has compact support, then the summation in (4.5.1) involves only a finite number of terms. The coefficients $\{p_k\}_k$ can be expressed in terms of the scaling function ϕ as follows. We multiply the equality (4.5.1) by $\sqrt{2}\phi(2x-l)$ and integrate with respect to x to obtain

$$p_l = \int_{\mathbb{R}} \sqrt{2}\, \phi(2x-l)\, \phi(x)\, dx, \quad l \in \mathbb{Z}. \tag{4.5.2}$$

Replacing x by $(2^j x - l)$ in (4.5.1), we derive the general scaling relation

$$\phi(2^j x - l) = \sum_k p_{k-2l} \sqrt{2}\, \phi(2^{j+1}x - k),$$

i.e.,

$$2^{j/2}\phi(2^j x - l) = \sum_k p_{k-2l} 2^{(j+1)/2} \phi(2^{j+1}x - k). \tag{4.5.3}$$

Corresponding to ϕ, define

$$\psi(x) = \sum_k (-1)^k p_{1-k} \sqrt{2}\, \phi(2x-k). \tag{4.5.4}$$

It can be shown that ψ is a wavelet function associated to the scaling function ϕ, and for any integer $j \in \mathbb{Z}$, the set $\{2^{j/2}\psi(2^j x - k) \mid k \in \mathbb{Z}\}$ is an orthonormal basis of the wavelet space W_j.

For a function ϕ satisfying (4.5.1) to be a scaling function of a multiresolution analysis, several identities must hold on the coefficients $\{p_k\}_k$, as a result of the orthonormality requirement (see Exercise 4.5.2 for a necessary condition). It is difficult to find the coefficients directly. One popular approach for the construction of the scaling and wavelet functions is through the Fourier transforms $\hat{\phi}$ and $\hat{\psi}$. Conditions on $\hat{\phi}$ and $\hat{\psi}$ can be identified for ϕ and ψ to be scaling and wavelet functions. This approach is especially useful for theoretical issues such as existence and smoothness properties of the scaling and wavelet functions. Detailed discussions of the subject can be found in [63], or in [38] for a more accessible presentation.

For functions defined through the scaling equation (4.5.1) (except in simple cases such as the Haar scaling function), we can not, in general, express them by closed formulas in terms of elementary functions. Nevertheless, in many cases, numerical values of $\phi(x)$ can be determined through a fixed-point iteration applied to the scaling equation (4.5.1): With a suitable starting function $\phi^{(0)}$, we compute a sequence of approximations $\{\phi^{(n)}\}$ by the recursion formula

$$\phi^{(n)}(x) = \sum_k p_k \sqrt{2}\, \phi^{(n-1)}(2x-k).$$

Numerical values of $\phi(x)$ can also be computed directly using the scaling equation (4.5.1), as is demonstrated in the next paragraph. Thus, the scaling and wavelet functions can be used for computations just like other classes of functions such as algebraic or trigonometric polynomials, exponential or logarithm functions, or special functions of mathematical physics.

As an example of Daubechies wavelets, consider a continuous scaling function $\phi(x)$ defined by the following properties:

$$\operatorname{supp} \phi = [0,3], \tag{4.5.5}$$

$$\begin{aligned}\phi(x) = {} & \frac{1+\sqrt{3}}{4}\,\phi(2x) + \frac{3+\sqrt{3}}{4}\,\phi(2x-1) \\ & + \frac{3-\sqrt{3}}{4}\,\phi(2x-2) + \frac{1-\sqrt{3}}{4}\,\phi(2x-3),\end{aligned} \tag{4.5.6}$$

$$\phi(1) = \frac{1+\sqrt{3}}{2}, \quad \phi(2) = \frac{1-\sqrt{3}}{2}. \tag{4.5.7}$$

Daubechies has shown that it is not possible to express $\phi(x)$ in terms of elementary mathematical functions through algebraic operations. Nevertheless, the properties (4.5.5)–(4.5.7) can be used to compute the values of

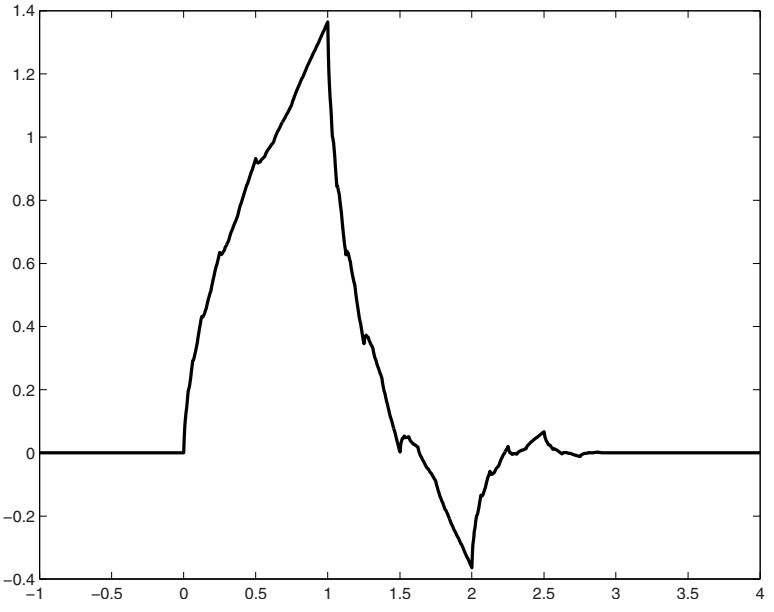

FIGURE 4.8. Daubechies scaling function

the function $\phi(x)$ at any point x. For example, we can compute $\phi(\frac{1}{2})$ as follows:

$$\begin{aligned}\phi\left(\frac{1}{2}\right) &= \frac{1+\sqrt{3}}{4}\,\phi(1) + \frac{3+\sqrt{3}}{4}\,\phi(0) + \frac{3-\sqrt{3}}{4}\,\phi(-1) + \frac{1-\sqrt{3}}{4}\,\phi(-2)\\ &= \frac{2+\sqrt{3}}{4}.\end{aligned}$$

The value of $\phi(x)$ for any $x \in \mathbb{D}_1$ (see page 193 for the definition of the set $\mathbb{D}_1$) can be computed similarly. Once the values of the function ϕ on $\mathbb{D}_1$ are known, we can continue to calculate the function values on $\mathbb{D}_2$, and so on.

A graph of the above Daubechies scaling function is shown in Figure 4.8.

Corresponding to the scaling function defined by (4.5.5)–(4.5.7), the wavelet function is

$$\begin{aligned}\psi(x) = &-\frac{1+\sqrt{3}}{4}\,\phi(2x-1) + \frac{3+\sqrt{3}}{4}\,\phi(2x)\\ &-\frac{3-\sqrt{3}}{4}\,\phi(2x+1) + \frac{1-\sqrt{3}}{4}\,\phi(2x+2).\end{aligned} \tag{4.5.8}$$

It is left as an exercise to verify that $\operatorname{supp}\psi = [-1, 2]$. A graph of this Daubechies wavelet function is shown in Figure 4.9.

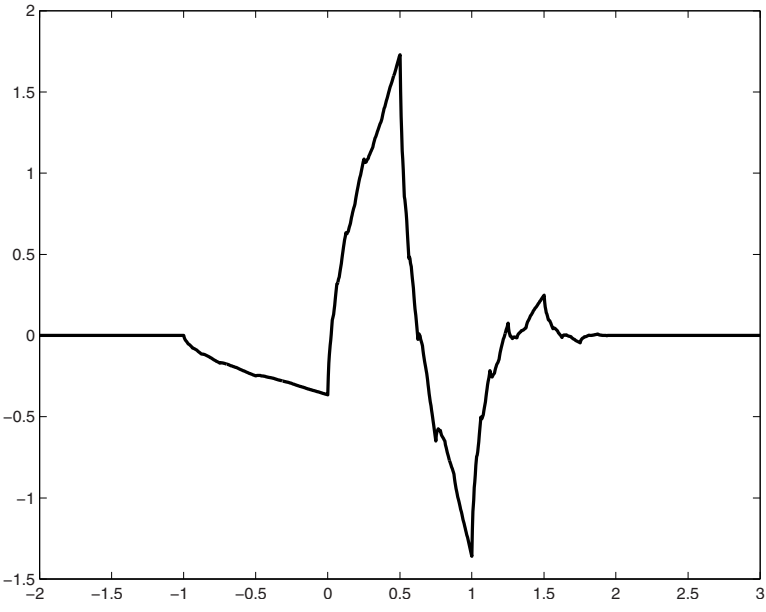

FIGURE 4.9. Daubechies wavelet function

The Daubechies wavelet function shown here is continuous, but not continuously differentiable. Wavelets with compact support and higher degree smoothness also exist.

Returning to the framework of a general multiresolution analysis, since $\{\psi_{jk}(x) \equiv 2^{j/2}\psi(2^j x - k)\}_{j,k\in\mathbb{Z}}$ is an orthonormal basis for $L^2(\mathbb{R})$, any $f \in L^2(\mathbb{R})$ can be expressed as

$$f(x) = \sum_j \sum_k b_{jk}\psi_{jk}(x).$$

The wavelet transform connects $f(x)$ to its wavelet coefficients $\{b_{jk}\}_{j,k\in\mathbb{Z}}$. Similar to the FFT for Fourier transforms, there is a recursive Fast Wavelet Transform with comparable speed and stability (see e.g., [217]). The decomposition and reconstruction procedures for the Haar system, as discussed in Section 4.4, can be extended to the other wavelets.

Exercise 4.5.1 Prove Proposition 4.5.2 using conditions (1) and (2) of Definition 4.5.1.

Exercise 4.5.2 Let ϕ satisfy the scaling equation (4.5.1). Show that a necessary condition for $\{\phi(x-k) \mid k \in \mathbb{Z}\}$ to be orthonormal is

$$\sum_k p_k p_{k-2l} = \delta_{0l}.$$

Exercise 4.5.3 Consider the space $L^2(-\pi, \pi)$. For each integer $j \geq 0$, let V_j be the subspace of trigonometric polynomials of degree less than or equal to j. Show that the conditions (3) and (4) (modified to $\overline{\cup_{j\in\mathbb{Z}} V_j} = L^2(-\pi, \pi)$) of Definition 4.5.1 are satisfied. Identify the subspace W_j in the orthogonal decomposition $V_{j+1} = V_j \oplus W_j$.

Exercise 4.5.4 For the Haar system discussed in Section 4.4, identify the coefficients in the scaling equation (4.5.1), and examine if the Haar wavelet function satisfies the relation (4.5.4).

Exercise 4.5.5 For the Daubechies scaling function defined in (4.5.5)–(4.5.7), verify the function values

$$\phi\left(\frac{3}{2}\right) = 0, \quad \phi\left(\frac{1}{4}\right) = \frac{5+3\sqrt{3}}{16}.$$

Exercise 4.5.6 For the Daubechies wavelet function given in (4.5.8), show that $\operatorname{supp}\psi = [-1, 2]$.
Hint: Recall (4.5.5).

Suggestion for Further Reading.

The classical reference to the Fourier series is Zygmund [251]. Deep mathematical theory of Fourier analysis can be found in the classic Stein and Weiss [212] or a more recent reference Duoandikoetxea [73]. Efficient implementation of the FFT is discussed in detail in Van Loan [226]. There, FFT is studied from the matrix factorization point of view, and the central theme of the book is the idea that different FFT algorithms correspond to different factorizations of the discrete Fourier transform matrix. Trefethen's monograph [223] provides practical Matlab implementations of the FFT and related algorithms, with applications in the numerical solution of ODEs and PDEs by the spectral method.

An elementary introduction of the theory of Fourier analysis and wavelets can be found in Boggess and Narcowich [38]. Two other good references on wavelets, without requiring much advanced mathematical background from the reader, are Kaiser [132] and Nievergelt [180]. Presentations of the deep mathematical theory of wavelets can be found in Bratteli and Jorgensen [41], Daubechies [63], Meyer [170], Wojtaszczyk [234]. The book by Strang and Nguyen [217] emphasizes the filter structures attached to wavelets, which are the key for their algorithmic efficiency and successful applications. Chui's book [51] emphasizes the important connection between wavelets and splines. A full exposition of wavelets and their generalizations, multiwaves, is found in Keinert [138]. Wavelets have been used in many branches of sciences and engineering, e.g., in computer graphics (Stollnita et al. [215]), geophysics (Foufoula-Georgiou and Kumar [83]), medicine and biology (Aldroubi and Unser [3]), signal processing (Mallat [164]). In this chapter, most of the discussion is confined

to the one-dimensional case. Discussion of multi-dimensional wavelets can be found in most of the references quoted above.

5

Nonlinear Equations and Their Solution by Iteration

Nonlinear functional analysis is the study of operators lacking the property of linearity. In this chapter, we consider nonlinear operator equations and their numerical solution. We begin the consideration of operator equations which take the form

$$u = T(u), \quad u \in K. \tag{5.0.1}$$

Here, K is a subset of a Banach space V, and $T : K \to V$. The solutions of this equation are called *fixed points* of the operator T, as they are left unchanged by T. The most important method for analyzing the solvability theory for such equations is the *Banach fixed-point theorem*. We present the Banach fixed-point theorem in Section 5.1 and then discuss its application to the study of various iterative methods in numerical analysis.

We then consider an extension of the well-known Newton method to the more general setting in Banach spaces. For this purpose, we introduce the differential calculus for nonlinear operators on normed spaces. The notion of the Gâteaux derivative leads to convenient characterizations for convexity of functionals and for minimizers of convex functionals.

We conclude the chapter with a brief introduction to another means of studying (5.0.1), using the concept of the *rotation of a completely continuous vector field*. We also generalize to function spaces the *conjugate gradient iteration method* for solving linear equations. There are many generalizations of the ideas of this chapter, and we intend this material as only a brief introduction.

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_5,
© Springer Science + Business Media, LLC 2009

5.1 The Banach fixed-point theorem

Let V be a Banach space with the norm $\|\cdot\|_V$, and let K be a subset of V. Consider an operator $T : K \to V$ defined on K. We are interested in the existence of a solution of the fixed-point problem (5.0.1) and the possibility of approximating the solution u by the following iterative method. Pick an initial guess $u_0 \in K$, and define a sequence $\{u_n\}$ by the iteration formula

$$u_{n+1} = T(u_n), \quad n = 0, 1, \ldots. \tag{5.1.1}$$

To have this make sense, we identify another requirement that must be imposed upon T:

$$T(v) \in K \quad \forall\, v \in K. \tag{5.1.2}$$

The problem of solving an equation

$$f(u) = 0 \tag{5.1.3}$$

for some operator $f : K \subset V \to V$ can be reduced to an equivalent fixed-point problem of the form (5.0.1) by setting $T(v) = v - c_0 f(v)$ for some constant scalar $c_0 \neq 0$, or more generally, $T(v) = v - F(f(v))$ with an operator $F : V \to V$ satisfying

$$F(w) = 0 \quad \Longleftrightarrow \quad w = 0.$$

Thus any result on the fixed-point problem (5.0.1) can be rephrased as a result for an equation (5.1.3). In addition, the iterative method (5.1.1) then provides a possible approximation procedure for solving the equation (5.1.3). In the following Section 5.2, we look at such applications for solving equations in a variety of settings.

For the iterative method to work, we must assume something more than (5.1.2). To build some insight as to what further assumptions are needed on the operator T, consider the following simple example.

Example 5.1.1 Take V to be the real line $\mathbb{R}$, and T an affine operator:

$$Tx = a\,x + b, \quad x \in \mathbb{R}$$

for some constants a and b. Now define the iterative method induced by the operator T. Let $x_0 \in \mathbb{R}$, and for $n = 0, 1, \ldots$, define

$$x_{n+1} = a\,x_n + b.$$

It is easy to see that

$$x_n = x_0 + n\,b \quad \text{if } a = 1,$$

and

$$x_n = a^n x_0 + \frac{1 - a^n}{1 - a}\,b \quad \text{if } a \neq 1.$$

Thus in the non-trivial case $a \neq 1$, the iterative method is convergent if and only if $|a| < 1$. Notice that the number $|a|$ occurs in the property

$$|Tx - Ty| \leq |a|\,|x - y| \quad \forall\, x, y \in \mathbb{R}.$$

□

Definition 5.1.2 *We say an operator $T : K \subset V \to V$ is* contractive *with contractivity constant $\alpha \in [0, 1)$ if*

$$\|T(u) - T(v)\|_V \leq \alpha\,\|u - v\|_V \quad \forall\, u, v \in K.$$

The operator T is called non-expansive *if*

$$\|T(u) - T(v)\|_V \leq \|u - v\|_V \quad \forall\, u, v \in K,$$

and Lipschitz continuous *if there exists a constant $L \geq 0$ such that*

$$\|T(u) - T(v)\|_V \leq L\,\|u - v\|_V \quad \forall\, u, v \in K.$$

We see the following implications:

$$\begin{array}{rcl} \text{contractivity} & \Longrightarrow & \text{non-expansiveness} \\ & \Longrightarrow & \text{Lipschitz continuity} \\ & \Longrightarrow & \text{continuity.} \end{array}$$

Theorem 5.1.3 (Banach fixed-point theorem) *Assume that K is a nonempty closed set in a Banach space V, and further, that $T : K \to K$ is a contractive mapping with contractivity constant α, $0 \leq \alpha < 1$. Then the following results hold.*

(1) *Existence and uniqueness: There exists a unique $u \in K$ such that*

$$u = T(u).$$

(2) *Convergence and error estimates of the iteration: For any $u_0 \in K$, the sequence $\{u_n\} \subset K$ defined by $u_{n+1} = T(u_n)$, $n = 0, 1, \ldots$, converges to u:*

$$\|u_n - u\|_V \to 0 \quad \text{as } n \to \infty.$$

For the error, the following bounds are valid:

$$\|u_n - u\|_V \leq \frac{\alpha^n}{1 - \alpha}\,\|u_0 - u_1\|_V, \tag{5.1.4}$$

$$\|u_n - u\|_V \leq \frac{\alpha}{1 - \alpha}\,\|u_{n-1} - u_n\|_V, \tag{5.1.5}$$

$$\|u_n - u\|_V \leq \alpha\,\|u_{n-1} - u\|_V. \tag{5.1.6}$$

Proof. Since $T : K \to K$, the sequence $\{u_n\}$ is well-defined. Let us first prove that $\{u_n\}$ is a Cauchy sequence. Using the contractivity of the mapping T, we have

$$\|u_{n+1} - u_n\|_V \le \alpha \|u_n - u_{n-1}\|_V \le \cdots \le \alpha^n \|u_1 - u_0\|_V.$$

Then for any $m \ge n \ge 1$,

$$\begin{aligned} \|u_m - u_n\|_V &\le \sum_{j=0}^{m-n-1} \|u_{n+j+1} - u_{n+j}\|_V \\ &\le \sum_{j=0}^{m-n-1} \alpha^{n+j} \|u_1 - u_0\|_V \\ &\le \frac{\alpha^n}{1-\alpha} \|u_1 - u_0\|_V. \end{aligned} \tag{5.1.7}$$

Since $\alpha \in [0, 1)$, $\|u_m - u_n\|_V \to 0$ as $m, n \to \infty$. Thus $\{u_n\}$ is a Cauchy sequence; and since K is a closed set in the Banach space V, $\{u_n\}$ has a limit $u \in K$. We take the limit $n \to \infty$ in $u_{n+1} = T(u_n)$ to see that $u = T(u)$ by the continuity of T, i.e., u is a fixed-point of T.

Suppose $u_1, u_2 \in K$ are both fixed-points of T. Then from $u_1 = T(u_1)$ and $u_2 = T(u_2)$, we obtain

$$u_1 - u_2 = T(u_1) - T(u_2).$$

Hence

$$\|u_1 - u_2\|_V = \|T(u_1) - T(u_2)\|_V \le \alpha \|u_1 - u_2\|_V$$

which implies $\|u_1 - u_2\|_V = 0$ since $\alpha \in [0, 1)$. So a fixed-point of a contractive mapping is unique.

Now we prove the error estimates. Letting $m \to \infty$ in (5.1.7), we get the estimate (5.1.4). From

$$\|u_n - u\|_V = \|T(u_{n-1}) - T(u)\|_V \le \alpha \|u_{n-1} - u\|_V$$

we obtain the estimate (5.1.6). This estimate together with

$$\|u_{n-1} - u\|_V \le \|u_{n-1} - u_n\|_V + \|u_n - u\|_V$$

implies the estimate (5.1.5). □

This theorem is called by a variety of names in the literature, with the *contractive mapping theorem* another popular choice. It is also called *Picard iteration* in settings related to differential equations.

As an application of the Banach fixed-point theorem, we consider the unique solvability of a nonlinear equation in a Hilbert space.

Theorem 5.1.4 *Let V be a Hilbert space. Assume $T : V \to V$ is strongly monotone and Lipschitz continuous, i.e., there exist two constants $c_1, c_2 > 0$ such that for any $v_1, v_2 \in V$,*

$$(T(v_1) - T(v_2), v_1 - v_2) \geq c_1 \|v_1 - v_2\|^2, \tag{5.1.8}$$

$$\|T(v_1) - T(v_2)\| \leq c_2 \|v_1 - v_2\|. \tag{5.1.9}$$

Then for any $b \in V$, there is a unique $u \in V$ such that

$$T(u) = b. \tag{5.1.10}$$

Moreover, the solution u depends Lipschitz continuously on b: If $T(u_1) = b_1$ and $T(u_2) = b_2$, then

$$\|u_1 - u_2\| \leq \frac{1}{c_1} \|b_1 - b_2\|. \tag{5.1.11}$$

Proof. The equation $T(u) = b$ is equivalent to

$$u = u - \theta \left[T(u) - b\right]$$

for any $\theta \neq 0$. Define an operator $T_\theta : V \to V$ by the formula

$$T_\theta(v) = v - \theta \left[T(v) - b\right].$$

Let us show that for $\theta > 0$ sufficiently small, the operator T_θ is contractive. Write

$$T_\theta(v_1) - T_\theta(v_2) = (v_1 - v_2) - \theta \left[T(v_1) - T(v_2)\right].$$

Then,

$$\begin{aligned} \|T_\theta(v_1) - T_\theta(v_2)\|^2 = \|v_1 - v_2\|^2 - 2\theta \left(T(v_1) - T(v_2), v_1 - v_2\right) \\ + \theta^2 \|T(v_1) - T(v_2)\|^2. \end{aligned}$$

Use the assumptions (5.1.8) and (5.1.9) to obtain

$$\|T_\theta(v_1) - T_\theta(v_2)\|^2 \leq \left(1 - 2c_2\theta + c_1^2\theta^2\right) \|v_1 - v_2\|^2.$$

For $\theta \in (0, 2c_2/c_1^2)$,

$$1 - 2c_2\theta + c_1^2\theta^2 < 1$$

and T_θ is a contraction. Then by the Banach fixed-point theorem, T_θ has a unique fixed-point $u \in V$. Hence, the equation (5.1.10) has a unique solution.

Now we prove the Lipschitz continuity of the solution with respect to the right hand side. From $T(u_1) = b_1$ and $T(u_2) = b_2$, we obtain

$$T(u_1) - T(u_2) = b_1 - b_2.$$

Then

$$(T(u_1) - T(u_2), u_1 - u_2) = (b_1 - b_2, u_1 - u_2)\,.$$

Apply the assumption (5.1.8) and the Cauchy-Schwarz inequality,

$$c_1\|u_1 - u_2\|^2 \leq \|b_1 - b_2\|\,\|u_1 - u_2\|,$$

which implies (5.1.11). □

The proof technique of Theorem 5.1.4 will be employed in Chapter 11 in proving existence and uniqueness of solutions to some variational inequalities. The condition (5.1.8) relates to the degree of monotonicity of $T(v)$ as v varies. For a real-valued function $T(v)$ of a single real variable v, the constant c_1 can be chosen as the infimum of $T'(v)$ over the domain of T, assuming this infimum is positive.

Exercise 5.1.1 In the Banach fixed-point theorem, we assume (1) V is a complete space, (2) K is a nonempty closed set in V, (3) $T : K \to K$, and (4) T is contractive. Find examples to show that each of these assumptions is necessary for the result of the theorem; in particular, the result fails to hold if all the other assumptions are kept except that the contractivity of T is replaced by the inequality

$$\|T(u) - T(v)\|_V < \|u - v\|_V \quad \forall\, u, v \in V,\ u \neq v.$$

Exercise 5.1.2 Assume K is a nonempty closed set in a Banach space V, and that $T : K \to K$ is continuous. Suppose T^m is a contraction for some positive integer m. Prove that T has a unique fixed-point in K. Moreover, prove that the iteration method

$$u_{n+1} = T(u_n), \quad n = 0, 1, 2, \ldots$$

converges.

Exercise 5.1.3 Let T be a contractive mapping on V to V. By Theorem 5.1.3, for every $y \in V$, the equation $v = T(v) + y$ has a unique solution, call it $u(y)$. Show that $u(y)$ is a continuous function of y.

Exercise 5.1.4 Let V be a Banach space, and let T be a contractive mapping on $K \subset V$ to K, with $K = \{v \in V \mid \|v\| \leq r\}$ for some $r > 0$. Assume $T(0) = 0$. Show that $v = T(v) + y$ has a unique solution in K for all sufficiently small choices of $y \in V$.

5.2 Applications to iterative methods

The Banach fixed-point theorem presented in the preceding section contains most of the desirable properties of a numerical method. Under the stated conditions, the approximation sequence is well-defined, and it is convergent to the unique solution of the problem. Furthermore, we know the

convergence rate is linear (see (5.1.6)), we have an *a priori* error estimate (5.1.4) which can be used to determine the number of iterations needed to achieve a prescribed solution accuracy before actual computations take place, and we also have an *a posteriori* error estimate (5.1.5) which gives a computable error bound once some numerical solutions are calculated.

In this section, we apply the Banach fixed-point theorem to the analysis of numerical approximations of several problems.

5.2.1 Nonlinear algebraic equations

Given a real-valued function of a real variable, $f : \mathbb{R} \to \mathbb{R}$, we are interested in computing its real roots, i.e., we are interested in solving the equation

$$f(x) = 0, \quad x \in \mathbb{R}. \tag{5.2.1}$$

There are a variety of ways to reformulate this equation as an equivalent fixed-point problem of the form

$$x = T(x), \quad x \in \mathbb{R}. \tag{5.2.2}$$

One example is $T(x) \equiv x - c_0 f(x)$ for some constant $c_0 \neq 0$. A more sophisticated example is $T(x) = x - f(x)/f'(x)$, in which case the iterative method becomes the celebrated Newton's method. For this last example, we generally use Newton's method only for finding simple roots of $f(x)$, which means we need to assume $f'(x) \neq 0$ when $f(x) = 0$. We return to a study of the Newton's method later in Section 5.4. Specializing the Banach fixed-point theorem to the problem (5.2.2), we have the following well-known result.

Theorem 5.2.1 *Let $-\infty < a < b < \infty$ and $T : [a, b] \to [a, b]$ be a contractive function with contractivity constant $\alpha \in [0, 1)$. Then the following results hold.*

(1) *Existence and uniqueness: There exists a unique solution $x \in [a, b]$ to the equation $x = T(x)$.*

(2) *Convergence and error estimates of the iteration: For any $x_0 \in [a, b]$, the sequence $\{x_n\} \subset [a, b]$ defined by $x_{n+1} = T(x_n)$, $n = 0, 1, \ldots$, converges to x:*

$$x_n \to x \quad \text{as } n \to \infty.$$

For the error, there hold the bounds

$$|x_n - x| \leq \frac{\alpha^n}{1-\alpha} |x_0 - x_1|,$$
$$|x_n - x| \leq \frac{\alpha}{1-\alpha} |x_{n-1} - x_n|,$$
$$|x_n - x| \leq \alpha |x_{n-1} - x|.$$

The contractiveness of the function T is guaranteed from the assumption that

$$\sup_{a\le x\le b} |T'(x)| < 1.$$

Indeed, using the Mean Value Theorem, we then see that T is contractive with the contractivity constant $\alpha = \sup_{a\le x\le b} |T'(x)|$.

5.2.2 Linear algebraic systems

Let $A \in \mathbb{R}^{m\times m}$ be an m by m matrix, and let us consider the linear system

$$A\boldsymbol{x} = \boldsymbol{b}, \quad \boldsymbol{x} \in \mathbb{R}^m \tag{5.2.3}$$

where $\boldsymbol{b} \in \mathbb{R}^m$ is given. It is well-known that (5.2.3) has a unique solution $\boldsymbol{x}$ for any given $\boldsymbol{b}$ if and only if A is non-singular, $\det(A) \neq 0$.

Let us reformulate (5.2.3) as a fixed point problem and introduce the corresponding iteration methods. A common practice for devising iterative methods to solve (5.2.3) is by using a matrix splitting

$$A = N - M$$

with N chosen in such a way that the system $N\boldsymbol{x} = \boldsymbol{k}$ is easily and uniquely solvable for any right side $\boldsymbol{k}$. Then the linear system (5.2.3) is rewritten as

$$N\boldsymbol{x} = M\boldsymbol{x} + \boldsymbol{b}.$$

This leads naturally to an iterative method for solving (5.2.3):

$$N\boldsymbol{x}_n = M\boldsymbol{x}_{n-1} + \boldsymbol{b}, \quad n = 1, 2, \ldots \tag{5.2.4}$$

with $\boldsymbol{x}_0$ a given initial guess of the solution $\boldsymbol{x}$.

To more easily analyze the iteration, we rewrite these last two equations as

$$\begin{aligned} \boldsymbol{x} &= N^{-1}M\boldsymbol{x} + N^{-1}\boldsymbol{b}, \\ \boldsymbol{x}_n &= N^{-1}M\boldsymbol{x}_{n-1} + N^{-1}\boldsymbol{b}. \end{aligned}$$

The matrix $N^{-1}M$ is called the iteration matrix. Subtracting the two equations, we obtain the error equation

$$\boldsymbol{x} - \boldsymbol{x}_n = N^{-1}M\left(\boldsymbol{x} - \boldsymbol{x}_{n-1}\right).$$

Inductively,

$$\boldsymbol{x} - \boldsymbol{x}_n = \left(N^{-1}M\right)^n \left(\boldsymbol{x} - \boldsymbol{x}_0\right), \quad n = 0, 1, 2, \ldots. \tag{5.2.5}$$

We see that the iterative method converges if $\|N^{-1}M\| < 1$, where $\|\cdot\|$ is some matrix operator norm, i.e., it is a norm induced by some vector norm $\|\cdot\|$:

$$\|A\| = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}. \tag{5.2.6}$$

Recall that for a square matrix A, a necessary and sufficient condition for $A^n \to 0$ as $n \to \infty$ is $r_\sigma(A) < 1$. This follows from the Jordan canonical form of a square matrix. Here, $r_\sigma(A)$ is the *spectral radius* of A:

$$r_\sigma(A) = \max_i |\lambda_i(A)|,$$

with $\{\lambda_i(A)\}$ the set of all the eigenvalues of A. Note that from the error relation (5.2.5), we have convergence $\boldsymbol{x}_n \to \boldsymbol{x}$ as $n \to \infty$ for any initial guess $\boldsymbol{x}_0$, if and only if $(N^{-1}M)^n \to 0$ as $n \to \infty$. Therefore, for the iterative method (5.2.4), a necessary and sufficient condition for convergence is $r_\sigma(N^{-1}M) < 1$.

The spectral radius of a matrix is an intrinsic quantity of the matrix, whereas a matrix norm is not. It is thus not surprising that a necessary and sufficient condition for convergence of the iterative method is described in terms of the spectral radius of the iteration matrix. We would also expect something of this kind based on the property that in finite dimensional spaces, convergence of $\{\boldsymbol{x}_n\}$ in one norm is equivalent to convergence in every other norm (see Theorem 1.2.14 from Chapter 1).

We have the following relations between the spectral radius and norms of a matrix $A \in \mathbb{R}^{m\times m}$.

1. $r_\sigma(A) \leq \|A\|$ for any matrix operator norm $\|\cdot\|$.
   This result follows immediately from the definition of $r_\sigma(A)$, the defining relation of an eigenvalue, and the fact that the matrix norm $\|\cdot\|$ is generated by a vector norm.

2. For any $\varepsilon > 0$, there exists a matrix operator norm $\|\cdot\|_{A,\varepsilon}$ such that

   $$r_\sigma(A) \leq \|A\|_{A,\varepsilon} \leq r_\sigma(A) + \varepsilon.$$

   For a proof, see [129, p. 12]. Thus,

   $$r_\sigma(A) = \inf\{\|A\| \mid \|\cdot\| \text{ is a matrix operator norm}\}.$$

3. $r_\sigma(A) = \lim_{n\to\infty} \|A^n\|^{1/n}$ for any matrix norm $\|\cdot\|$.
   Note that here the norm can be any matrix norm, not necessarily the ones generated by vector norms as in (5.2.6). This can be proven by using the Jordan canonical form; see [15, p. 490].

For applications to the solution of discretizations of Laplace's equation and some other elliptic partial differential equations, it is useful to write

$$A = D + L + U,$$

where D is the diagonal part of A, L and U are the strict lower and upper triangular parts. If we take $N = D$, then (5.2.4) reduces to

$$D\,\boldsymbol{x}_n = \boldsymbol{b} - (L+U)\,\boldsymbol{x}_{n-1},$$

which is the vector representation of the Jacobi method; the corresponding componentwise representation is

$$x_{n,i} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_{n-1,j}\right), \quad 1 \le i \le m.$$

If we take $N = D + L$, then we obtain the Gauss-Seidel method

$$(D+L)\,\boldsymbol{x}_n = \boldsymbol{b} - U\,\boldsymbol{x}_{n-1}$$

or equivalently,

$$x_{n,i} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_{n,j} - \sum_{j=i+1}^{m} a_{ij}x_{n-1,j}\right), \quad 1 \le i \le m.$$

A more sophisticated splitting is obtained by setting

$$N = \frac{1}{\omega}\,D + L, \qquad M = \frac{1-\omega}{\omega}\,D - U,$$

where $\omega \neq 0$ is an acceleration parameter. The corresponding iterative method with the (approximate) optimal choice of ω is called the SOR (successive overrelaxation) method. The componentwise representation of the SOR method is

$$x_{n,i} = (1-\omega)\,x_{n-1,i} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_{n,j} - \sum_{j=i+1}^{m} a_{ij}x_{n-1,j}\right),\ 1 \le i \le m.$$

For linear systems arising in difference solutions of some model partial differential equation problems, there is a well-understood theory for the choice of an optimal value of ω; and with that optimal value, the iteration converges much more rapidly than does the original Gauss-Seidel method on which it is based. Additional discussion of the framework (5.2.4) for iteration methods is given in [15, Section 8.6].

5.2.3 Linear and nonlinear integral equations

Recall Example 2.3.2 from Chapter 2, in which we discussed solvability of the integral equation

$$\lambda u(x) - \int_a^b k(x,y)\,u(y)\,dy = f(x), \quad a \le x \le b \tag{5.2.7}$$

by means of the geometric series theorem. For simplicity, we assume $k \in C([a,b] \times [a,b])$ and let $f \in C[a,b]$, although these assumptions can be weakened considerably. In Example 2.3.2, we established that within the framework of the function space $C[a,b]$ with the uniform norm, the equation (5.2.7) was uniquely solvable if

$$\max_{a \le x \le b} \int_a^b |k(x,y)|\, dy < |\lambda|. \tag{5.2.8}$$

If we rewrite the equation (5.2.7) as

$$u(x) = \frac{1}{\lambda} \int_a^b k(x,y)\, u(y)\, dy + \frac{1}{\lambda} f(x), \quad a \le x \le b$$

which has the form $u = T(u)$, then we can apply the Banach fixed-point theorem. Doing so, it is straightforward to derive a formula for the contractivity constant:

$$\alpha = \frac{1}{|\lambda|} \max_{a \le x \le b} \int_a^b |k(x,y)|\, dy.$$

The requirement that $\alpha < 1$ is exactly the assumption (5.2.8). Moreover, the fixed point iteration

$$u_n(x) = \frac{1}{\lambda} \int_a^b k(x,y)\, u_{n-1}(y)\, dy + \frac{1}{\lambda} f(x), \quad a \le x \le b, \tag{5.2.9}$$

for $n = 1, 2, \ldots$, can be shown to be equivalent to a truncation of the geometric series for solving (5.2.7). This is left as Exercise 5.2.5.

Nonlinear integral equations of the second kind

Nonlinear integral equations lack the property of linearity. Consequently, we must assume other properties in order to be able to develop a solvability theory for them. We discuss here some commonly seen nonlinear integral equations of the second kind. The integral equation

$$u(x) = \mu \int_a^b k(x,y,u(y))\, dy + f(x), \quad a \le x \le b \tag{5.2.10}$$

is called a *Urysohn integral equation*. Here we assume that

$$f \in C[a,b] \quad \text{and} \quad k \in C([a,b] \times [a,b] \times \mathbb{R}). \tag{5.2.11}$$

Moreover, we assume k satisfies a *uniform Lipschitz condition* with respect to its third argument:

$$|k(x,y,u_1) - k(x,y,u_2)| \le M\, |u_1 - u_2|, \quad a \le x, y \le b,\ u_1, u_2 \in \mathbb{R}\,. \tag{5.2.12}$$

Since (5.2.10) is of the form $v = T(v)$, we can introduce the fixed point iteration

$$u_n(x) = \mu \int_a^b k(x, y, u_{n-1}(y))\, dy + f(x), \quad a \le x \le b,\ n \ge 1. \tag{5.2.13}$$

Theorem 5.2.2 *Assume f and k satisfy the conditions* (5.2.11), (5.2.12). *Moreover, assume*

$$|\mu|\, M\, (b-a) < 1.$$

Then the integral equation (5.2.10) *has a unique solution $u \in C[a, b]$, and it can be approximated by the iteration method of* (5.2.13).

Another well-studied nonlinear integral equation is

$$u(x) = \mu \int_a^b k(x, y)\, h(y, u(y))\, dy + f(x), \quad a \le x \le b$$

with $k(x, y), h(y, u)$, and $f(x)$ given. This is called a *Hammerstein integral equation.* These equations are often derived as reformulations of boundary value problems for nonlinear ordinary differential equations. Multi-variate generalizations of this equation are obtained as reformulations of boundary value problems for nonlinear elliptic partial differential equations.

An interesting nonlinear integral equation which does not fall into the above categories is *Nekrasov's equation:*

$$\theta(x) = \lambda \int_0^\pi L(x, t) \frac{\sin\theta(t)}{1 + 3\lambda \displaystyle\int_0^t \sin\theta(s)\, ds}\, dt, \quad 0 \le x \le \pi, \tag{5.2.14}$$

where

$$L(x, t) = \frac{1}{\pi} \log \frac{\sin((x+t)/2)}{\sin((x-t)/2)}.$$

One solution is $\theta(x) \equiv 0$, and it is the nonzero solutions that are of interest. This arises in the study of the profile of water waves on liquids of infinite depth; and the equation involves interesting questions of solutions that bifurcate. See [173, p. 415].

Nonlinear Volterra integral equations of the second kind

An equation of the form

$$u(t) = \int_a^t k(t, s, u(s))\, ds + f(t), \quad t \in [a, b] \tag{5.2.15}$$

is called a nonlinear Volterra integral equation of the second kind. When $k(t, s, u)$ depends linearly on u, we get a linear Volterra integral equation,

and such equations were investigated earlier in Example 2.3.4 of Section 2.3. The form of the equation (5.2.15) leads naturally to the iterative method

$$u_n(t) = \int_a^t k(t, s, u_{n-1}(s))\, ds + f(t), \quad t \in [a, b],\ n \geq 1. \tag{5.2.16}$$

Theorem 5.2.3 *Assume $k(t, s, u)$ is continuous for $a \leq s \leq t \leq b$ and $u \in \mathbb{R}$; and let $f \in C[a, b]$. Furthermore, assume*

$$|k(t, s, u_1) - k(t, s, u_2)| \leq M\,|u_1 - u_2|\,, \quad a \leq s \leq t \leq b,\ u_1, u_2 \in \mathbb{R}$$

for some constant M. Then the integral equation (5.2.15) *has a unique solution $u \in C[a, b]$. Moreover, the iterative method* (5.2.16) *converges for any initial function $u_0 \in C[a, b]$.*

Proof. There are at least two approaches to applying the Banach fixed-point theorem to prove the existence of a unique solution of (5.2.15). We give a sketch of the two approaches below, assuming the conditions stated in Theorem 5.2.3. We define the nonlinear integral operator

$$T : C[a, b] \to C[a, b], \quad Tu(t) \equiv \int_a^t k(t, s, u(s))\, ds + f(t).$$

Approach 1. Let us show that for m sufficiently large, the operator T^m is a contraction on $C[a, b]$. For $u, v \in C[a, b]$,

$$Tu(t) - Tv(t) = \int_a^t [k(t, s, u(s)) - k(t, s, v(s))]\, ds.$$

Then

$$|Tu(t) - Tv(t)| \leq M \int_a^t |u(s) - v(s)|\, ds \tag{5.2.17}$$

and

$$|Tu(t) - Tv(t)| \leq M\|u - v\|_\infty (t - a).$$

Since

$$T^2u(t) - T^2v(t) = \int_a^t [k(t, s, Tu(s)) - k(t, s, Tv(s))]\, ds,$$

we get

$$\begin{aligned} \left|T^2u(t) - T^2v(t)\right| &\leq M \int_a^t |Tu(s) - Tv(s)|\, ds \\ &\leq \frac{[M(t - a)]^2}{2!}\, \|u - v\|_\infty. \end{aligned}$$

By a mathematical induction, we obtain

$$|T^m u(t) - T^m v(t)| \le \frac{[M(t-a)]^m}{m!} \|u - v\|_\infty.$$

Thus

$$\|T^m u - T^m v\|_\infty \le \frac{[M(b-a)]^m}{m!} \|u - v\|_\infty.$$

Since

$$\frac{[M(b-a)]^m}{m!} \to 0 \quad \text{as } m \to \infty,$$

the operator T^m is a contraction on $C[a,b]$ when m is chosen sufficiently large. By the result in Exercise 5.1.2, the operator T has a unique fixed-point in $C[a,b]$ and the iteration sequence converges to the solution. Derivation of error bounds is left as an exercise.

Approach 2. Over the space $C[a,b]$, let us introduce the norm

$$|||v||| = \max_{a \le t \le b} e^{-\beta t} |v(t)|$$

which is equivalent to the standard norm $\|v\|_\infty$ on $C[a,b]$. The parameter β is chosen to satisfy $\beta > M$. We then modify the relation (5.2.17) as follows:

$$e^{-\beta t} |Tu(t) - Tv(t)| \le M e^{-\beta t} \int_a^t e^{\beta s} e^{-\beta s} |u(s) - v(s)| \, ds.$$

Hence,

$$\begin{aligned} e^{-\beta t} |Tu(t) - Tv(t)| &\le M e^{-\beta t} |||u - v||| \int_a^t e^{\beta s} ds \\ &= \frac{M}{\beta} e^{-\beta t} (e^{\beta t} - e^{\beta a}) \, |||u - v|||. \end{aligned}$$

Therefore,

$$|||Tu - Tv||| \le \frac{M}{\beta} |||u - v|||.$$

Since $\beta > M$, the operator T is a contraction on the Banach space $(V, |||\cdot|||)$. Then T has a unique fixed-point which is the unique solution of the integral equation (5.2.15) and the iteration sequence converges. □

We observe that if the stated assumptions are valid over the interval $[a, \infty)$, then the conclusions of Theorem 5.2.3 remain true on $[a, \infty)$. This implies that the equation

$$u(t) = \int_a^t k(t, s, u(s)) \, ds + f(t), \quad t \ge a$$

has a unique solution $u \in C[a,\infty)$; and for any $b > a$, we have the convergence $\|u - u_n\|_{C[a,b]} \to 0$ as $n \to \infty$ with $\{u_n\} \subset C[a,\infty)$ being defined by

$$u_n(t) = \int_a^t k(t,s,u_{n-1}(s))\,ds + f(t), \quad t \geq a.$$

Note that although the value of the Lipschitz constant M may increase as $(b-a)$ increases, the result will remain valid.

5.2.4 Ordinary differential equations in Banach spaces

Let V be a Banach space and consider the initial value problem

$$\begin{cases} u'(t) = f(t,u(t)), & |t-t_0| < a, \\ u(t_0) = z. \end{cases} \tag{5.2.18}$$

Here $z \in V$ and $f : [t_0 - a, t_0 + a] \times V \to V$ is continuous. For example, f could be an integral operator; and then (5.2.18) would be an "integro-differential equation". The differential equation problem (5.2.18) is equivalent to the integral equation

$$u(t) = z + \int_{t_0}^t f(s,u(s))\,ds, \quad |t-t_0| < a, \tag{5.2.19}$$

which is of the form $u = T(u)$. This leads naturally to the fixed point iteration method

$$u_n(t) = z + \int_{t_0}^t f(s,u_{n-1}(s))\,ds, \quad |t-t_0| < a,\ n \geq 1. \tag{5.2.20}$$

Denote, for $b > 0$,

$$Q_b \equiv \{(t,u) \in \mathbb{R} \times V \mid |t-t_0| \leq a,\ \|u-z\| \leq b\}.$$

We have the following existence and solvability theory for (5.2.18). The proof is a straightforward application of Theorem 5.1.3 and the ideas incorporated in the proof of Theorem 5.2.3.

Theorem 5.2.4 (Generalized Picard-Lindelöf Theorem) *Assume $f : Q_b \to V$ is continuous and is uniformly Lipschitz continuous with respect to its second argument:*

$$\|f(t,u) - f(t,v)\| \leq L\,\|u-v\| \quad \forall\,(t,u),(t,v) \in Q_b,$$

where L is a constant independent of t. Let

$$M = \max_{(t,u)\in Q_b} \|f(t,u)\|$$

and

$$a_0 = \min\left\{a, \frac{b}{M}\right\}.$$

Then the initial value problem (5.2.18) *has a unique continuously differentiable solution* $u(\cdot)$ *on* $[t_0 - a_0, t_0 + a_0]$*; and the iterative method* (5.2.20) *converges for any initial value* u_0 *for which* $\|z - u_0\| < b$*,*

$$\max_{|t-t_0|\le a_0} \|u_n(t) - u(t)\| \to 0 \quad \text{as } n \to \infty.$$

Moreover, with $\alpha = 1 - e^{-L\,a_0}$*, the error*

$$\max_{|t-t_0|\le a_0} \|u_n(t) - u(t)\| e^{-L\,|t-t_0|}$$

is bounded by each of the following:

$$\frac{\alpha^n}{1-\alpha} \max_{|t-t_0|\le a_0} \|u_1(t) - u_0(t)\| e^{-L\,|t-t_0|},$$
$$\frac{\alpha}{1-\alpha} \max_{|t-t_0|\le a_0} \|u_{n-1}(t) - u_n(t)\| e^{-L\,|t-t_0|},$$
$$\alpha \max_{|t-t_0|\le a_0} \|u_{n-1}(t) - u(t)\| e^{-L\,|t-t_0|}.$$

Exercise 5.2.1 This exercise illustrates the effect of the reformulation of the equation on the convergence of the iterative method. As an example, we compute the positive square root of 2, which is a root of the equation $x^2 - 2 = 0$. First reformulating the equation as $x = 2/x$, we obtain an iterative method $x_n = 2/x_{n-1}$. Show that unless $x_0 = \sqrt{2}$, the method is not convergent.
(*Hint*: Compare x_{n+1} with x_{n-1}.)

Then let us consider another reformulation. Notice that $\sqrt{2} \in [1, 2]$ and is a fixed-point of the equation

$$x = T(x) \equiv \frac{1}{4}(2 - x^2) + x.$$

Verify that $T : [1, 2] \to [1, 2]$ and $\max_{1\le x\le 2} |T'(x)| = 1/2$. Thus with any $x_0 \in [1, 2]$, the iterative method

$$x_n = \frac{1}{4}(2 - x_{n-1}^2) + x_{n-1}, \quad n \ge 1$$

is convergent.

Exercise 5.2.2 A matrix $A = (a_{ij})$ is called *diagonally dominant* if

$$\sum_{j\ne i} |a_{ij}| < |a_{ii}| \quad \forall\, i.$$

Apply the Banach fixed-point theorem to show that if A is diagonally dominant, then both the Jacobi method and the Gauss-Seidel method converge.

Exercise 5.2.3 A simple iteration method can be developed for the linear system (5.2.3) as follows. For a parameter $\theta \neq 0$, write the system in the equivalent form

$$\boldsymbol{x} = \boldsymbol{x} + \theta\,(\boldsymbol{b} - A\,\boldsymbol{x}),$$

and introduce the iteration formula

$$\boldsymbol{x}_n = \boldsymbol{x}_{n-1} + \theta\,(\boldsymbol{b} - A\,\boldsymbol{x}_{n-1}).$$

Assume A is symmetric and positive definite, and denote its largest eigenvalue by $\lambda_{\max}$. Show that for any initial guess $\boldsymbol{x}_0$, the iteration method converges if and only if

$$0 < \theta < \frac{2}{\lambda_{\max}}.$$

Determine the optimal value of θ so that the spectral radius of the iteration matrix is minimal.

Exercise 5.2.4 Prove Theorem 5.2.2. In addition, state error bounds for the iteration (5.2.13) based on (5.1.4)–(5.1.6).

Exercise 5.2.5 Show that the iteration (5.2.9) is equivalent to some truncation of the geometric series for (5.2.7). Apply the fixed point theorem to derive error bounds for the iteration based on (5.1.4)–(5.1.6).

Exercise 5.2.6 Derive error bounds for the iteration of Theorem 5.2.3.

Exercise 5.2.7 Let $f \in C[0,1]$ be given. Consider the following integral equation on $C[0,1]$:

$$u(x) = \mu \int_0^1 \sqrt{x^2 + u(y)^2}\,dy + f(x), \quad 0 \leq x \leq 1.$$

Give a bound on μ that guarantees a unique solution.

Do the same for the integral equation:

$$u(x) = \mu \int_0^x \sqrt{x^2 + u(y)^2}\,dy + f(x), \quad 0 \leq x \leq 1.$$

Exercise 5.2.8 Consider the Fredholm integral equation

$$u(t) = c_0 \int_0^1 \sin(t\,u(s))\,ds + f(t),$$

where $c_0 \in \mathbb{R}$ and $f \in C[0,1]$. Determine a range of c_0 for which the integral equation admits a unique solution.

Exercise 5.2.9 Generalize Theorem 5.2.3 to a system of d Volterra integral equations. Specifically, consider the equation

$$\boldsymbol{u}(t) = \int_a^t \boldsymbol{k}(t,s,\boldsymbol{u}(s))\,ds + \boldsymbol{f}(t), \quad t \in [a,b].$$

In this equation, $\boldsymbol{u}(t), \boldsymbol{f}(t) \in \mathbb{R}^d$ and $\boldsymbol{k}(t,s,\boldsymbol{u})$ is an $\mathbb{R}^d$ valued function of $a \leq s \leq t \leq b$ and $\boldsymbol{u} \in \mathbb{R}$. Include error bounds for the corresponding iterative method.

Exercise 5.2.10 Apply Theorem 5.2.3 to show that the initial value problem

$$u'' + p(x)\,u' + q(x)\,u = f(x), \quad x > 0,$$
$$u(0) = u_0, \quad u'(0) = v_0$$

has a unique solution $u \in C^2[0,\infty)$. Here, $u_0, v_0 \in \mathbb{R}$ and $p, q, f \in C[0,\infty)$ are given.
Hint: Convert the initial value problem to a Volterra integral equation of the second kind for u''.

Exercise 5.2.11 Prove the generalized Picard-Lindelöf theorem.

Exercise 5.2.12 Gronwall's inequality provides an upper bound for a continuous function f on $[a,b]$ which satisfies the relation

$$f(t) \le g(t) + \int_a^t h(s)\,f(s)\,ds, \quad t \in [a,b],$$

where g is continuous, $h \in L^1(a,b)$, and $h(t) \ge 0$ a.e. Show that

$$f(t) \le g(t) + \int_a^t g(s)\,h(s) \exp\Big(\int_s^t h(\tau)\,d\tau\Big)\,ds \quad \forall\, t \in [a,b].$$

Moreover if g is nondecreasing, then

$$f(t) \le g(t) \exp\Big(\int_a^t h(s)\,ds\Big) \quad \forall\, t \in [a,b].$$

In the special case when $h(s) = c > 0$, these inequalities reduce to

$$f(t) \le g(t) + c\int_a^t g(s)\,e^{c\,(t-s)}ds \quad \forall\, t \in [a,b]$$

and

$$f(t) \le g(t)\,e^{c\,(t-a)} \quad \forall\, t \in [a,b],$$

respectively.

Exercise 5.2.13 Gronwall's inequality is useful in stability analysis. Let $f : [t_0 - a, t_0 + a] \times V \to V$ be continuous and Lipschitz continuous with respect to its second argument:

$$\|f(t,u) - f(t,v)\| \le L\,\|u - v\| \quad \forall\, t \in [t_0 - a, t_0 + a],\ u, v \in V.$$

Let $u_{1,0}, u_{2,0} \in V$, and let r_1 and r_2 be continuous mappings from $[t_0 - a, t_0 + a]$ to V. Define u_1 and u_2 by

$$u_1'(t) = f(t, u_1(t)) + r_1(t), \quad u_1(t_0) = u_{1,0},$$
$$u_2'(t) = f(t, u_2(t)) + r_2(t), \quad u_2(t_0) = u_{2,0}.$$

Show that

$$\|u_1(t) - u_2(t)\| \le e^{L\,|t-t_0|}\left[\|u_{1,0} - u_{2,0}\| + a \max_{|s-t_0| \le |t-t_0|} \|r_1(s) - r_2(s)\|\right].$$

Thus, the solution of the differential equation depends continuously on the source term r and the initial value.

Gronwall's inequality and its discrete analog are useful also in error estimates of some numerical methods.

5.3 Differential calculus for nonlinear operators

In this section, we generalize the notion of derivatives of real functions to that of operators. General references for this material are [33, Section 2.1], [135, Chap. 17].

5.3.1 Fréchet and Gâteaux derivatives

We first recall the definition of the derivative of a real function. Let I be an interval on $\mathbb{R}$, and x_0 an interior point of I. A function $f : I \to \mathbb{R}$ is differentiable at x_0 if and only if

$$f'(x_0) \equiv \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \text{ exists;} \tag{5.3.1}$$

or equivalently, for some number a,

$$f(x_0 + h) = f(x_0) + a\,h + o(|h|) \text{ as } h \to 0. \tag{5.3.2}$$

where we let $f'(x_0) = a$ denote the derivative.

From the eyes of a first year calculus student, for a real-valued real-variable function, the definition (5.3.1) looks simpler than (5.3.2), though the two definitions are equivalent. Nevertheless, the definition (5.3.2) clearly indicates that the nature of differentiation is (local) linearization. Moreover, the form (5.3.2) can be directly extended to define the derivative of a general operator, whereas the form (5.3.1) is useful for defining directional or partial derivatives of the operator. We illustrate this by looking at a vector-valued function of several real variables.

Let K be a subset of the space $\mathbb{R}^d$, with $\boldsymbol{x}_0$ as an interior point. Let $\boldsymbol{f} : K \to \mathbb{R}^m$. Following (5.3.2), we say $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}_0$ if there exists a matrix (linear operator) $A \in \mathbb{R}^{m \times d}$ such that

$$\boldsymbol{f}(\boldsymbol{x}_0 + \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}_0) + A\boldsymbol{h} + o(\|\boldsymbol{h}\|) \quad \text{as } \boldsymbol{h} \to \boldsymbol{0},\ \boldsymbol{h} \in \mathbb{R}^d. \tag{5.3.3}$$

We can show that $A = \nabla \boldsymbol{f}(\boldsymbol{x}_0) = (A_{ij})$, the gradient or Jacobian of $\boldsymbol{f}$ at $\boldsymbol{x}_0$:

$$A_{ij} = \frac{\partial f_i}{\partial x_j}, \quad 1 \le i \le m,\ 1 \le j \le d.$$

There is a difficulty in extending (5.3.1) for the differentiability of multi-variable functions: how to extend the meaning of the divided difference $[f(x_0 + h) - f(x_0)]/h$ when h is a vector? On the other hand, (5.3.1) can be extended directly to provide the notion of a directional derivative: We do not linearize the function in all the possible directions of the variable $\boldsymbol{x}$ approaching $\boldsymbol{x}_0$; rather, we linearize the function along a fixed direction towards $\boldsymbol{x}_0$. In this way, we will only need to deal with a vector-valued function of one real variable, and then the divided difference in (5.3.1)

makes sense. More precisely, let $\boldsymbol{h}$ be a fixed vector in $\mathbb{R}^d$, and we consider the function $\boldsymbol{f}(\boldsymbol{x}_0 + t\,\boldsymbol{h})$ for $t \in \mathbb{R}$ in a neighborhood of 0. We then say $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}_0$ with respect to $\boldsymbol{h}$, if there is a matrix A such that

$$\lim_{t\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}_0 + t\,\boldsymbol{h}) - \boldsymbol{f}(\boldsymbol{x}_0)}{t} = A\boldsymbol{h}. \tag{5.3.4}$$

In case $\|\boldsymbol{h}\| = 1$, we call the quantity $A\boldsymbol{h}$ the directional derivative of $\boldsymbol{f}$ at $\boldsymbol{x}_0$ along the direction $\boldsymbol{h}$. We notice that if $\boldsymbol{f}$ is differentiable at $\boldsymbol{x}_0$ following the definition (5.3.3), then (5.3.4) is also valid. But the converse is not true: The relation (5.3.4) for any $\boldsymbol{h} \in \mathbb{R}^d$ does not imply the relation (5.3.3); see Exercise 5.3.2.

We now turn to the case of an operator $f : K \subset V \to W$ between two normed spaces V and W. Let us adopt the convention that whenever we discuss the differentiability at a point u_0, implicitly we assume u_0 is an interior point of K; by this, we mean there is an $r > 0$ such that

$$B(u_0, r) \equiv \{u \in V \mid \|u - u_0\| \le r\} \subset K.$$

Definition 5.3.1 *The operator f is* Fréchet differentiable *at u_0 if and only if there exists $A \in \mathcal{L}(V, W)$ such that*

$$f(u_0 + h) = f(u_0) + Ah + o(\|h\|), \quad h \to 0. \tag{5.3.5}$$

The map A is called the Fréchet derivative of f at u_0, and we write $A = f'(u_0)$. The quantity $df(u_0; h) = f'(u_0)h$ is called the Fréchet differential *of f at u_0 along h. If f is Fréchet differentiable at all points in $K_0 \subset K$, we say f is Fréchet differentiable on K_0 and call $f' : K_0 \subset V \to \mathcal{L}(V, W)$ the* Fréchet derivative *of f on K_0.*

If f is differentiable at u_0, then the derivative $f'(u_0)$ is unique. This is verified as follows. Suppose there exists another map $\tilde{A} \in \mathcal{L}(V, W)$ such that

$$f(u_0 + h) = f(u_0) + \tilde{A}h + o(\|h\|), \quad h \to 0.$$

Then

$$\tilde{A}h - f'(u_0)h = o(\|h\|), \quad h \to 0.$$

For any $h_0 \in V$ with $\|h_0\| = 1$, let $h = t\,h_0$, $0 \ne t \in \mathbb{R}$. Dividing the relation by t and taking the limit $t \to 0$, we obtain

$$\tilde{A}h_0 - f'(u_0)h_0 = 0.$$

Hence, $\tilde{A} = f'(u_0)$.

Definition 5.3.2 *The operator f is* Gâteaux differentiable *at u_0 if and only if there exists $A \in \mathcal{L}(V, W)$ such that*

$$\lim_{t\to 0} \frac{f(u_0 + t\,h) - f(u_0)}{t} = Ah \quad \forall\, h \in V. \tag{5.3.6}$$

The map A is called the Gâteaux derivative *of f at u_0, and we write $A = f'(u_0)$. The quantity $df(u_0; h) = f'(u_0)h$ is called the* Gâteaux differential *of f at u_0 along h. If f is Gâteaux differentiable at all points in $K_0 \subset K$, we say f is Gâteaux differentiable on K_0 and call $f' : K_0 \subset V \to \mathcal{L}(V, W)$ the Gâteaux derivative of f on K_0.*

From the defining relation (5.3.5), we immediately obtain the next result.

Proposition 5.3.3 *If $f'(u_0)$ exists as a Fréchet derivative, then f is continuous at u_0.*

Evidently, the relation (5.3.6) is equivalent to

$$f(u_0 + t\,h) = f(u_0) + t\,Ah + o(|t|) \quad \forall\, h \in V.$$

Thus a Fréchet derivative is also the Gâteaux derivative. The converse of this statement is not true, as is shown in Exercise 5.3.2. Nevertheless, we have the following result.

Proposition 5.3.4 *A Fréchet derivative is also a Gâteaux derivative. Conversely, if the limit in* (5.3.6) *is uniform with respect to h with $\|h\| = 1$ or if the Gâteaux derivative is continuous at u_0, then the Gâteaux derivative at u_0 is also the Fréchet derivative at u_0.*

Now we present some differentiation rules. If we do not specify the type of derivative, then the result is valid for both the Fréchet derivative and the Gâteaux derivative.

Proposition 5.3.5 (Sum rule) *Let V and W be normed spaces. If $f, g : K \subset V \to W$ are differentiable at u_0, then for any scalars α and β, $\alpha f + \beta g$ is differentiable at u_0 and*

$$(\alpha f + \beta g)'(u_0) = \alpha f'(u_0) + \beta g'(u_0).$$

Proposition 5.3.6 (Product rule) *Let V, V_1, V_2 and W be normed spaces. If $f_1 : K \subset V \to V_1$ and $f_2 : K \subset V \to V_2$ are differentiable at u_0, and $b : V_1 \times V_2 \to W$ is a bounded bilinear form, then the operator $B(u) = b(f_1(u), f_2(u))$ is differentiable at u_0, and*

$$B'(u_0)h = b(f_1'(u_0)h, f_2(u_0)) + b(f_1(u_0), f_2'(u_0)h), \quad h \in V.$$

Proposition 5.3.7 (Chain rule) *Let U, V and W be normed spaces. Let $f : K \subset U \to V$, $g : L \subset V \to W$ be given with $f(K) \subset L$. Assume u_0 is an interior point of K, $f(u_0)$ is an interior point of L. If $f'(u_0)$ and $g'(f(u_0))$ exist as Fréchet derivatives, then $g \circ f$ is Fréchet differentiable at u_0 and*

$$(g \circ f)'(u_0) = g'(f(u_0))f'(u_0).$$

If $f'(u_0)$ exists as a Gâteaux derivative and $g'(f(u_0))$ exists as a Fréchet derivative, then $g \circ f$ is Gâteaux differentiable at u_0 and the above formula holds.

Let us look at some examples.

Example 5.3.8 Let $f : V \to W$ be a continuous affine operator,

$$f(v) = Lv + b,$$

where $L \in \mathcal{L}(V, W)$, $b \in W$, and $v \in V$. Then f is Fréchet differentiable, and $f'(v) = L$ is constant. □

Example 5.3.9 For a function $T : K \subset \mathbb{R}^m \to \mathbb{R}^n$, the Fréchet derivative is the $n \times m$ Jacobian matrix evaluated at $v_0 = (x_1, \ldots, x_m)^T$:

$$T'(v_0) = \left(\frac{\partial T_i(v_0)}{\partial x_j} \right)_{\substack{1 \le i \le n \\ 1 \le j \le m}}.$$

□

Example 5.3.10 Let $V = W = C[a, b]$ with the maximum norm. Assume $g \in C[a, b]$, $k \in C([a, b] \times [a, b] \times \mathbb{R})$. Then we can define the operator $T : V \to W$ by the formula

$$T(u)(t) = g(t) + \int_a^b k(t, s, u(s))\, ds.$$

The integral operator in this formula is called a *Urysohn integral operator*. Let $u_0 \in C[a, b]$ be such that

$$\frac{\partial k}{\partial u}(t, s, u_0(s)) \in C([a, b]^2).$$

Then T is Fréchet differentiable at u_0, and

$$(T'(u_0)h)(t) = \int_a^b \frac{\partial k}{\partial u}(t, s, u_0(s))\, h(s)\, ds, \quad h \in V.$$

The restriction that $k \in C([a, b] \times [a, b] \times \mathbb{R})$ can be relaxed in a number of ways, with the definition of $T'(u_0)$ still valid. □

It is possible to introduce Fréchet and Gâteaux derivatives of higher order. For example, the second Fréchet derivative is the derivative of the Fréchet derivative. For $f : K \subset V \to W$ differentiable on $K_0 \subset K$, the Fréchet derivative is a mapping $f' : K_0 \subset V \to W$. If f' is Fréchet differentiable on K_0, then the second Fréchet derivative

$$f'' = (f')' : K_0 \subset V \to \mathcal{L}(V, \mathcal{L}(V, W)).$$

At each point $v \in K_0$, the second derivative $f''(v)$ can also be viewed as a bilinear mapping from $V \times V$ to W, and

$$f'' : K_0 \subset V \to \mathcal{L}(V \times V, W),$$

and this is generally the way f'' is regarded. Detailed discussions on Fréchet and Gâteaux derivatives, including higher order derivatives, are given in [135, Section 17.2] and [244, Section 4.5].

5.3.2 Mean value theorems

Let us generalize the mean-value theorem for differentiable functions of a real variable. This then allows us to consider the effect on a nonlinear function of perturbations in its argument.

Proposition 5.3.11 *Let U and V be real Banach spaces, and let $F : K \subset U \to V$ with K an open set. Assume F is differentiable on K and that $F'(u)$ is a continuous function of u on K to $\mathcal{L}(U, V)$. Let $u, w \in K$ and assume the line segment joining them is also contained in K. Then*

$$\|F(u) - F(w)\|_V \leq \sup_{0 \leq \theta \leq 1} \|F'((1-\theta)u + \theta w)\| \, \|u - w\|_U. \tag{5.3.7}$$

Proof. Denote $y = F(u) - F(w)$. The Hahn-Banach theorem in the form of Corollary 2.5.6 justifies the existence of a linear functional $T : V \to \mathbb{R}$ with $\|T\| = 1$ and $T(y) = \|y\|_V$. Introduce the real-valued function

$$g(t) = T(F(tu + (1-t)\, w)), \quad 0 \leq t \leq 1.$$

Note that $T(y) = g(1) - g(0)$.

We show g is continuously differentiable on $[0, 1]$ using the chain rule of Proposition 5.3.7. Introduce

$$\begin{aligned} g_1(t) &= tu + (1-t)\, w, \quad & g_1 &: [0,1] \to V, \\ g_2(v) &= T(F(v)), & g_2 &: K \subset V \to \mathbb{R}. \end{aligned}$$

For $0 \leq t \leq 1$,

$$\begin{aligned} g(t) &= g_2(g_1(t)), \\ g'(t) &= g_2'(g_1(t)) g_1'(t) \\ &= [T \circ F'(tu + (1-t)\, w)]\,(u - w) \\ &= T\left[F'(tu + (1-t)\, w)\,(u - w)\right]. \end{aligned}$$

Applying the ordinary mean-value theorem, we have a $\theta \in [0, 1]$ for which

$$\begin{aligned} \|F(u) - F(w)\|_V &= g(1) - g(0) = g'(\theta) \\ &= T\left[F'(\theta u + (1-\theta)\, w)\,(u - w)\right] \\ &\leq \|T\| \, \|F'(\theta u + (1-\theta)\, w)\,(u - w)\|_W \\ &\leq \|F'(\theta u + (1-\theta)\, w)\| \, \|u - w\|_U. \end{aligned}$$

The inequality (5.3.7) follows immediately. □

Corollary 5.3.12 *Let U and V be normed spaces, and let K be a connected open set in U. Assume $F : K \to V$ is differentiable. If $F'(v) = 0$ for any $v \in K$, then F is a constant function.*

For a continuously differentiable function $F : \mathbb{R} \to \mathbb{R}$, (5.3.7) follows from the ordinary mean-value theorem

$$F(u) - F(w) = F'((1-\theta)\,u + \theta\,w)\,(u-w), \quad u, w \in \mathbb{R},$$

for some $\theta \in [0,1]$ depending on F, u and w. However, this form of the mean-value theorem does not hold for functions defined on general Banach spaces; see Exercises 5.3.7 and 5.3.8.

The following result provides an error bound for the linear Taylor approximation for a nonlinear function. A proof similar to the above can be given for this lemma.

Proposition 5.3.13 *Let U and V be real Banach spaces, and let $F : K \subset U \to V$ with K an open set. Assume F is twice continuously differentiable on K, with $F'' : K \to \mathcal{L}(U \times U, V)$. Let $u_0, u_0 + h \in K$ along with the line segment joining them. Then*

$$\|F(u_0+h) - [F(u_0) + F'(u_0)h]\|_V \le \frac{1}{2} \sup_{0\le\theta\le 1} \|F''(u_0+\theta h)\|\, \|h\|_U^2.$$

5.3.3 Partial derivatives

The following definition is given for either type of derivatives (Fréchet or Gâteaux).

Definition 5.3.14 *Let U, V and W be Banach spaces, $f : \mathcal{D}(f) \subset U \times V \to W$. For fixed $v_0 \in V$, $f(u, v_0)$ is a function of u whose derivative at u_0, if it exists, is called the partial derivative of f with respect to u, and is denoted by $f_u(u_0, v_0)$. The partial derivative $f_v(u_0, v_0)$ is defined similarly.*

We explore the relation between the Fréchet derivative and partial Fréchet derivatives.

Proposition 5.3.15 *If f is Fréchet differentiable at (u_0, v_0), then the partial Fréchet derivatives $f_u(u_0, v_0)$ and $f_v(u_0, v_0)$ exist, and*

$$f'(u_0, v_0)(h, k) = f_u(u_0, v_0)\,h + f_v(u_0, v_0)\,k, \quad h \in U,\ k \in V. \tag{5.3.8}$$

Conversely, if $f_u(u, v)$ and $f_v(u, v)$ exist in a neighborhood of (u_0, v_0) and are continuous at (u_0, v_0), then f is Fréchet differentiable at (u_0, v_0), and (5.3.8) *holds.*

Proof. Assume f is Fréchet differentiable at (u_0, v_0), then

$$f(u_0 + h, v_0 + k) = f(u_0, v_0) + f'(u_0, v_0)(h, k) + o(\|(h, k)\|).$$

Setting $k = 0$, we obtain

$$f(u_0 + h, v_0) = f(u_0, v_0) + f'(u_0, v_0)(h, 0) + o(\|h\|).$$

Therefore, $f_u(u_0, v_0)$ exists and

$$f_u(u_0, v_0)\, h = f'(u_0, v_0)(h, 0).$$

Similarly, $f_v(u_0, v_0)$ exists and

$$f_v(u_0, v_0)\, k = f'(u_0, v_0)(0, k).$$

Adding the two relations, we get (5.3.8).

Now assume $f_u(u, v)$ and $f_v(u, v)$ exist in a neighborhood of (u_0, v_0) and are continuous at (u_0, v_0). We have

$$\begin{aligned}
&\|f(u_0 + h, v_0 + k) - [f(u_0, v_0) + f_u(u_0, v_0)\, h + f_v(u_0, v_0)\, k]\| \\
&\quad \le \|f(u_0 + h, v_0 + k) - [f(u_0, v_0 + k) + f_u(u_0, v_0 + k)\, h]\,\| \\
&\qquad + \|f_u(u_0, v_0 + k)\, h - f_u(u_0, v_0)\, h\| \\
&\qquad + \|f(u_0, v_0 + k) - [f(u_0, v_0) + f_v(u_0, v_0)\, k]\| \\
&\quad = o(\|(h, k)\|).
\end{aligned}$$

Hence, f is Fréchet differentiable at (u_0, v_0). □

Existence of partial derivatives at a point does not imply the existence of the Fréchet or Gâteaux derivative (Exercise 5.3.1).

Corollary 5.3.16 *A mapping $f(u, v)$ is continuously Fréchet differentiable in a neighborhood of (u_0, v_0) if and only if $f_u(u, v)$ and $f_v(u, v)$ are continuous in a neighborhood of (u_0, v_0).*

The above discussion can be extended straightforward to maps of several variables.

5.3.4 The Gâteaux derivative and convex minimization

Let us first use the notion of Gâteaux derivative to characterize the convexity of Gâteaux differentiable functionals.

Theorem 5.3.17 *Let V be a normed space and $K \subset V$ be a non-empty convex subset. Assume $f : K \to \mathbb{R}$ is Gâteaux differentiable. Then the following three statements are equivalent.*
(a) *f is convex.*
(b) *$f(v) \ge f(u) + \langle f'(u), v - u\rangle \ \forall\, u, v \in K$.*
(c) *$\langle f'(v) - f'(u), v - u\rangle \ge 0 \ \forall\, u, v \in K$.*

Proof. (a) $\Longrightarrow$ (b). For any $t \in [0,1]$, by the convexity of f,

$$f(u + t\,(v-u)) \le t\,f(v) + (1-t)\,f(u).$$

Then

$$\frac{f(u + t\,(v-u)) - f(u)}{t} \le f(v) - f(u), \quad t \in (0,1].$$

Taking the limit $t \to 0+$, we obtain

$$\langle f'(u), v-u\rangle \le f(v) - f(u).$$

(b) $\Longrightarrow$ (a). For any $u, v \in K$, any $\lambda \in [0,1]$, we have

$$f(v) \ge f(u + \lambda\,(v-u)) + (1-\lambda)\,\langle f'(u + \lambda\,(v-u)), v-u\rangle,$$
$$f(u) \ge f(u + \lambda\,(v-u)) - \lambda\,\langle f'(u + \lambda\,(v-u)), v-u\rangle.$$

Multiplying the first inequality by λ, the second inequality by $1-\lambda$, and adding the two relations, we obtain

$$\lambda\,f(v) + (1-\lambda)\,f(u) \ge f(u + \lambda\,(v-u)).$$

So f is a convex function.
(b) $\Longrightarrow$ (c). For any $u, v \in K$, we have

$$f(v) \ge f(u) + \langle f'(u), v-u\rangle,$$
$$f(u) \ge f(v) + \langle f'(v), u-v\rangle.$$

Add the two inequalities to obtain

$$0 \ge -\langle f'(v) - f'(u), v-u\rangle,$$

i.e. (c) holds.
(c) $\Longrightarrow$ (b). Define a real function

$$\phi(t) = f(u + t\,(v-u)), \quad t \in [0,1].$$

Using Taylor's theorem, we have

$$\phi(1) = \phi(0) + \phi'(\theta) \quad \text{for some } \theta \in (0,1).$$

Notice that

$$\phi(1) = f(v), \quad \phi(0) = f(u).$$

Also

$$\begin{aligned}\phi'(\theta) &= \langle f'(u + \theta\,(v-u)), v-u\rangle \\ &= \frac{1}{\theta}\,\langle f'(u + \theta\,(v-u)) - f'(u), \theta\,(v-u)\rangle + \langle f'(u), v-u\rangle \\ &\ge \langle f'(u), v-u\rangle,\end{aligned}$$

where in the last step we used the condition (c). Thus (b) holds. □

There is a similar result on the strict convexity of a Gâteaux differentiable functional. Its proof is left as an exercise.

Theorem 5.3.18 *Let V be a normed space and $K \subset V$ be a non-empty convex subset. Assume $f : K \to \mathbb{R}$ is Gâteaux differentiable. Then the following three statements are equivalent.*
(a) *f is strictly convex.*
(b) *$f(v) > f(u) + \langle f'(u), v - u\rangle \ \forall\, u, v \in K,\ u \neq v$.*
(c) *$\langle f'(v) - f'(u), v - u\rangle > 0 \ \forall\, u, v \in K,\ u \neq v$.*

Let us then characterize minimizers of Gâteaux differentiable convex functionals.

Theorem 5.3.19 *Let V be a normed space and $K \subset V$ be a non-empty convex subset. Assume $f : K \to \mathbb{R}$ is convex and Gâteaux differentiable. Then there exists $u \in K$ such that*

$$f(u) = \inf_{v \in K} f(v) \tag{5.3.9}$$

if and only if there exists $u \in K$ such that

$$\langle f'(u), v - u\rangle \geq 0 \quad \forall\, v \in K. \tag{5.3.10}$$

When K is a subspace, the inequality (5.3.10) *reduces to an equality:*

$$\langle f'(u), v\rangle = 0 \quad \forall\, v \in K. \tag{5.3.11}$$

Proof. Assume u satisfies (5.3.9). Then for any $t \in (0, 1)$, since $u + t\,(v - u) \in K$, we have

$$f(u) \leq f(u + t\,(v - u)).$$

Then we have (5.3.10) by an argument similar to the one used in the proof of Theorem 5.3.17 for the part "(a) $\Longrightarrow$ (b)".

Now assume u satisfies (5.3.10). Then since f is convex,

$$f(v) \geq f(u) + \langle f'(u), v - u\rangle \geq f(u).$$

When K is a subspace, we can take v in (5.3.10) to be $v + u$ for any $v \in K$ to obtain

$$\langle f'(u), v\rangle \geq 0 \quad \forall\, v \in K.$$

Since K is a subspace, $-v \in K$ and

$$\langle f'(u), -v\rangle \geq 0 \quad \forall\, v \in K.$$

Therefore, we have the equality (5.3.11). □

Exercise 5.3.1 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x_1, x_2) = \begin{cases} \dfrac{x_1 x_2^2}{x_1^2 + x_2^4}, & \text{if } (x_1, x_2) \neq (0, 0), \\ 0, & \text{if } (x_1, x_2) = (0, 0). \end{cases}$$

Show that at $(0, 0)$, the two partial derivatives $f_{x_1}(0, 0)$ and $f_{x_2}(0, 0)$ exist, however, f is neither Gâteaux nor Fréchet differentiable at $(0, 0)$.

Exercise 5.3.2 Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$f(x_1, x_2) = \begin{cases} \dfrac{x_1 x_2^3}{x_1^2 + x_2^4}, & \text{if } (x_1, x_2) \neq (0, 0), \\ 0, & \text{if } (x_1, x_2) = (0, 0). \end{cases}$$

Show that at $(0, 0)$, the function is Gâteaux differentiable and is not Fréchet differentiable.

Exercise 5.3.3 Let V be an inner product space, and define $f(v) = \|v\|^2$ for $v \in V$. Compute $f'(u)$ at any $u \in V$.

Discuss the differentiability of the norm function $g(v) = \|v\|$, $v \in V$.

Exercise 5.3.4 Prove Proposition 5.3.5.

Exercise 5.3.5 Prove Proposition 5.3.6.

Exercise 5.3.6 Prove Proposition 5.3.7.

Exercise 5.3.7 Define a function $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^2$ by the formula

$$\boldsymbol{f}(x) = (\sin x, \cos x)^T, \quad x \in \mathbb{R}.$$

Compute its Fréchet derivative $\boldsymbol{f}'(x)$. Show that there is no θ such that

$$\boldsymbol{f}(2\pi) - \boldsymbol{f}(0) = \boldsymbol{f}'((1-\theta)\,0 + \theta\,2\pi)\,(2\pi - 0).$$

Exercise 5.3.8 Assume $\boldsymbol{f} : \mathbb{R} \to \mathbb{R}^m$ is continuously differentiable. Show the integral form of the mean-value theorem:

$$\boldsymbol{f}(b) - \boldsymbol{f}(a) = \int_0^1 \boldsymbol{f}'((1-t)\,a + t\,b)\,dt\,(b - a).$$

In particular, this formula holds for the function in Exercise 5.3.7.

Exercise 5.3.9 Give an example to show that the connectedness of the set K in Corollary 5.3.12 cannot be dropped.

Exercise 5.3.10 Prove Proposition 5.3.13.

Exercise 5.3.11 Following the proof of Theorem 5.3.17, give a proof of Theorem 5.3.18.

Hint: In proving (a) $\Longrightarrow$ (b), use the inequality

$$\langle f'(u), v - u \rangle \leq \frac{f(u + t\,(v-u)) - f(u)}{t}, \quad t \in (0, 1].$$

Exercise 5.3.12 Convexity of functionals can also be characterized by using second derivatives. In this exercise, we restrict our consideration to real functions of vector variables; though the results can be extended to functionals defined on general normed spaces. Let $K \subset \mathbb{R}^d$ be non-empty and convex, $f : K \to \mathbb{R}$ be twice continuously Gâteaux differentiable. Recall that the matrix

$$\nabla^2 f(\boldsymbol{x}) = \left(\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} \right)_{d \times d}$$

is usually called the Hessian matrix of f.
(a) Show that f is convex on K if and only if the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is positive semi-definite for any $\boldsymbol{x} \in K$.
(b) If the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ is positive definite for any $\boldsymbol{x} \in K$, then f is strictly convex on K.

Exercise 5.3.13 Show that for $p \geq 1/2$, the real-valued function

$$f(\boldsymbol{\xi}) = \frac{1}{p} \left(1 + |\boldsymbol{\xi}|^2\right)^p, \quad \boldsymbol{\xi} \in \mathbb{R}^d$$

is strictly convex.
Hint: Apply the result from Exercise 5.3.12 (b).

Exercise 5.3.14 Let $f : C^1[0,1] \to C[0,1]$ be defined by

$$f(u) = \left(\frac{du}{dx} \right)^2, \quad u \in C^1[0,1].$$

Calculate $f'(u)$.

Exercise 5.3.15 Let $A \in \mathcal{L}(V)$ be self-adjoint, V being a real Hilbert space. Define

$$f(v) = \frac{1}{2} (Av, v).$$

Then $f : V \to \mathbb{R}$. Show the existence of the Fréchet derivative $f'(v)$ and calculate it. Do the same for $A \in \mathcal{L}(V)$ without being necessarily self-adjoint.

Exercise 5.3.16 Let V be a real Hilbert space, $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ a symmetric, continuous bilinear form, and $\ell \in V'$. Compute the Fréchet derivative of the functional

$$f(v) = \frac{1}{2} a(v, v) - \ell(v), \quad v \in V.$$

Exercise 5.3.17 (a) Find the derivative of the nonlinear operator given in the right hand side of (5.2.14).
(b) Let $u(t) = \sin \theta(t)$, and reformulate (5.2.14) as a new fixed point problem $u = K(u)$. Find $K'(u)$ for $u = 0$.

5.4 Newton's method

Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable and consider the equation

$$f(x) = 0.$$

Suppose we know an approximate solution x_n near a root of the equation x^*. Then by the Taylor expansion,

$$\begin{aligned} 0 &= f(x^*) \\ &= f(x_n) + f'(x_n)\,(x^* - x_n) + o(|x^* - x_n|) \\ &\approx f(x_n) + f'(x_n)\,(x^* - x_n). \end{aligned}$$

Thus,

$$x^* \approx x_n - [f'(x_n)]^{-1} f(x_n).$$

This leads to the well-known Newton method for solving the equation $f(x) = 0$:

$$x_{n+1} = x_n - [f'(x_n)]^{-1} f(x_n), \quad n = 0, 1, \ldots.$$

In this section, we generalize the Newton method to solving operators equations.

5.4.1 Newton's method in Banach spaces

Let U and V be two Banach spaces. Assume $F : U \to V$ is Fréchet differentiable. We are interested in solving the equation

$$F(u) = 0. \tag{5.4.1}$$

The Newton method reads as follows: Choose an initial guess $u_0 \in U$; for $n = 0, 1, \ldots$, compute

$$u_{n+1} = u_n - [F'(u_n)]^{-1} F(u_n). \tag{5.4.2}$$

Theorem 5.4.1 (Local convergence) *Assume u^* is a solution of the equation (5.4.1) such that $[F'(u^*)]^{-1}$ exists and is a continuous linear map from V to U. Assume further that $F'(u)$ is locally Lipschitz continuous at u^*,*

$$\|F'(u) - F'(v)\| \le L\,\|u - v\| \quad \forall\, u, v \in N(u^*),$$

where $N(u^*)$ is a neighborhood of u^*, and $L > 0$ is a constant. Then there exists a $\delta > 0$ such that if $\|u_0 - u^*\| \le \delta$, the Newton's sequence $\{u_n\}$ is well-defined and converges to u^*. Furthermore, for some constant M with $M\,\delta < 1$, we have the error bounds

$$\|u_{n+1} - u^*\| \le M\,\|u_n - u^*\|^2 \tag{5.4.3}$$

and

$$\|u_n - u^*\| \le (M\delta)^{2^n}/M. \tag{5.4.4}$$

Proof. Upon redefining the neighborhood $N(u^*)$ if necessary, we may assume $[F'(u)]^{-1}$ exists on $N(u^*)$ and

$$c_0 = \sup_{u \in N(u^*)} \|[F'(u)]^{-1}\| < \infty.$$

Let us define

$$T(u) = u - [F'(u)]^{-1} F(u), \quad u \in N(u^*).$$

Notice that $T(u^*) = u^*$. For $u \in N(u^*)$, we have

$$\begin{aligned} T(u) - T(u^*) &= u - u^* - [F'(u)]^{-1} F(u) \\ &= [F'(u)]^{-1} \left[F(u^*) - F(u) - F'(u)\,(u^* - u) \right] \\ &= [F'(u)]^{-1} \int_0^1 [F'(u + t\,(u^* - u)) - F'(u)]\, dt\, (u^* - u) \end{aligned}$$

and by taking the norm,

$$\begin{aligned} \|T(u) - T(u^*)\| &\le \|[F'(u)]^{-1}\| \int_0^1 \|F'(u + t\,(u^* - u)) - F'(u)\|\, dt\, \|u^* - u\| \\ &\le \|[F'(u)]^{-1}\| \int_0^1 L\, t\, \|u^* - u\|\, dt\, \|u^* - u\|. \end{aligned}$$

Hence,

$$\|T(u) - T(u^*)\| \le \frac{c_0 L}{2} \|u - u^*\|^2. \tag{5.4.5}$$

Choose $\delta < 2/(c_0 L)$ with the property $\overline{B}(u^*, \delta) \subset N(u^*)$; and note that

$$\alpha \equiv c_0 L \delta / 2 < 1.$$

Then (5.4.5) implies

$$\begin{aligned} \|T(u) - u^*\| &= \|T(u) - T(u^*)\| \\ &\le \alpha \|u - u^*\|, \qquad u \in \overline{B}(u^*, \delta). \end{aligned} \tag{5.4.6}$$

Assume an initial guess $u_0 \in \overline{B}(u^*, \delta)$. Then (5.4.6) implies

$$\|u_1 - u^*\| = \|T(u_0) - u^*\| \le \alpha \|u_0 - u^*\| \le \alpha\delta < \delta.$$

Thus $u_1 \in \overline{B}(u^*, \delta)$. Repeating this argument inductively, we have $u_n \in \overline{B}(u^*, \delta)$ for all $n \ge 0$.

To obtain the convergence of $\{u_n\}$, we begin with

$$\|u_{n+1} - u^*\| = \|T(u_n) - u^*\| \le \alpha \|u_n - u^*\|, \quad n \ge 0.$$

By induction,

$$\|u_n - u^*\| \le \alpha^n \|u_0 - u^*\|, \quad n \ge 0$$

and $u_n \to u^*$ as $n \to \infty$.

Returning to (5.4.5), denote $M = c_0 L/2$. Note that $M\,\delta = \alpha < 1$. Then we get the estimate

$$\|u_{n+1} - u^*\| = \|T(u_n) - u^*\| \le M\,\|u_n - u^*\|^2,$$

proving (5.4.3). Multiply both sides by M, obtaining

$$M\|u_{n+1} - u^*\| \le (M\,\|u_n - u^*\|)^2\,.$$

An inductive application of this inequality leads to

$$M\,\|u_n - u^*\| \le (M\,\|u_0 - u^*\|)^{2^n},$$

thus proving (5.4.4). □

The theorem clearly shows that the Newton method is locally convergent with quadratic convergence. The main drawback of the result is the dependence of the assumptions on a root of the equation, which is the quantity to be computed. The Kantorovich theory overcomes this difficulty. A proof of the following theorem can be found in [244, p. 210].

Theorem 5.4.2 (KANTOROVICH) *Suppose that*
(a) $F : \mathcal{D}(F) \subset U \to V$ *is differentiable on an open convex set* $\mathcal{D}(F)$*, and the derivative is Lipschitz continuous:*

$$\|F'(u) - F'(v)\| \le L\,\|u - v\| \quad \forall\, u, v \in \mathcal{D}(F).$$

(b) *For some* $u_0 \in \mathcal{D}(F)$, $[F'(u_0)]^{-1}$ *exists and is a continuous operator from* V *to* U*, and such that* $h = a\,b\,L \le 1/2$ *for some* $a \ge \|[F'(u_0)]^{-1}\|$ *and* $b \ge \|[F'(u_0)]^{-1}F(u_0)\|$*. Denote*

$$t^* = \frac{1 - (1-2h)^{1/2}}{a\,L}, \quad t^{**} = \frac{1 + (1-2h)^{1/2}}{a\,L}.$$

(c) u_0 *is chosen so that* $\overline{B}(u_1, r) \subset \mathcal{D}(F)$*, where* $r = t^* - b$.
Then the equation (5.4.1) *has a solution* $u^* \in \overline{B}(u_1, r)$ *and the solution is unique in* $\overline{B}(u_0, t^{**}) \cap \mathcal{D}(F)$*; the sequence* $\{u_n\}$ *converges to* u^**, and we have the error estimate*

$$\|u_n - u^*\| \le \frac{\left[1 - (1-2h)^{1/2}\right]^{2^n}}{2^n a\,L}, \quad n = 0, 1, \ldots.$$

The Kantorovich theorem provides sufficient conditions for the convergence of the Newton method. These conditions are usually difficult to verify. Nevertheless, at least theoretically, the result is of great importance. For other related discussions of Newton's method, see [33, pp. 116–118] and [135, Chap. 18].

5.4.2 Applications

Nonlinear systems

Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a continuously differentiable function. A nonlinear system is of the form

$$\boldsymbol{x} \in \mathbb{R}^d, \quad F(\boldsymbol{x}) = \mathbf{0}.$$

Then the Newton method is

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - [F'(\boldsymbol{x}_n)]^{-1} F(\boldsymbol{x}_n), \tag{5.4.7}$$

which can also be written in the form

$$F'(\boldsymbol{x}_n)\, \boldsymbol{\delta}_n = -F(\boldsymbol{x}_n), \quad \boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \boldsymbol{\delta}_n.$$

So at each step, we solve a linear system. The method breaks down when $F'(\boldsymbol{x}_n)$ is singular or nearly singular.

Nonlinear integral equations

Consider the nonlinear integral equation

$$u(t) = \int_0^1 k(t, s, u(s))\, ds \tag{5.4.8}$$

over the space $U = C[0,1]$. Assume $k \in C([0,1] \times [0,1] \times \mathbb{R})$ and is continuously differentiable with respect to its third argument. Introducing an operator $F : U \to U$ through the formula

$$F(u)(t) = u(t) - \int_0^1 k(t, s, u(s))\, ds, \quad t \in [0,1],$$

the integral equation can be written in the form

$$F(u) = 0.$$

Newton's method for the problem is

$$u_{n+1} = u_n - [F'(u_n)]^{-1} F(u_n),$$

or equivalently,

$$F'(u_n)(u_{n+1} - u_n) = -F(u_n). \tag{5.4.9}$$

Let us compute the derivative of F.

$$\begin{aligned} F'(u)(v)(t) &= \lim_{h \to 0} \frac{1}{h} \left[F(u + hv)(t) - F(u)(t) \right] \\ &= \lim_{h \to 0} \frac{1}{h} \left\{ h\, v(t) - \int_0^1 \left[k(t, s, u(s) + h\, v(s)) - k(t, s, u(s)) \right] ds \right\} \\ &= v(t) + \int_0^1 \frac{\partial k(t, s, u(s))}{\partial u}\, v(s)\, ds. \end{aligned}$$

Therefore, the Newton iteration formula is

$$\begin{aligned}\delta_{n+1}(t) - \int_0^1 \frac{\partial k(t,s,u_n(s))}{\partial u}\,\delta_{n+1}(s)\,ds \\ = -u_n(t) + \int_0^1 k(t,s,u_n(s))\,u_n(s)\,ds, \qquad (5.4.10)\\ u_{n+1}(t) = u_n(t) + \delta_{n+1}(t).\end{aligned}$$

At each step, we solve a linear integral equation.

It is often computationally more efficient to use a modification of (5.4.9), using a fixed value of the derivative:

$$F'(u_0)(u_{n+1} - u_n) = -F(u_n). \qquad (5.4.11)$$

The iteration formula is now

$$\begin{aligned}\delta_{n+1}(t) - \int_0^1 \frac{\partial k(t,s,u_0(s))}{\partial u}\,\delta_{n+1}(s)\,ds \\ = -u_n(t) + \int_0^1 k(t,s,u_n(s))\,u_n(s)\,ds, \qquad (5.4.12)\\ u_{n+1}(t) = u_n(t) + \delta_{n+1}(t).\end{aligned}$$

This converges more slowly; but the lack of change in the integral equation (since only the right side is varying) often leads to less computation than with (5.4.10).

Nonlinear differential equations

As a sample problem, we consider

$$\begin{cases} u''(t) = f(t,u(t)), \quad t \in (0,1), \\ u(0) = u(1) = 0. \end{cases}$$

Here the function $f : [0,1]\times\mathbb{R}$ is assumed to be continuous and continuously differentiable with respect to its second argument. We take

$$U = C_0^2[0,1] = \left\{v \in C^2[0,1] \mid v(0) = v(1) = 0\right\}$$

with the norm $\|\cdot\|_{C^2[0,1]}$. Define

$$F(u)(t) = u''(t) - f(t,u(t)), \quad t \in [0,1].$$

It can be shown that

$$F'(u)(y)(t) = y''(t) - \frac{\partial f(t,u(t))}{\partial u}\,y(t).$$

Thus at each step, we solve a linearized boundary value problem

$$\begin{cases} u''_{n+1}(t) - \dfrac{\partial f}{\partial u}(t,u_n(t))\,u_{n+1}(t) = f(t,u_n(t)) - \dfrac{\partial f}{\partial u}(t,u_n(t))\,u_n(t), \\ \qquad\qquad t \in (0,1), \\ u_{n+1}(0) = u_{n+1}(1) = 0. \end{cases}$$

Exercise 5.4.1 Give the Newton method for solving the nonlinear boundary value problem

$$\begin{aligned} &-u''(x) + u(x)\,u'(x) + u(x)^3 = e^x, \quad x \in (0,1), \\ &u(0) = 1, \quad u'(1) = 2. \end{aligned}$$

Exercise 5.4.2 Consider the nonlinear advection-diffusion problem

$$\begin{aligned} \alpha \boldsymbol{u} - \mu\,\Delta \boldsymbol{u} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} &= \boldsymbol{f} \quad \text{in } \Omega, \\ \boldsymbol{u} &= \boldsymbol{g}_0 \quad \text{on } \Gamma_0, \\ \mu\,\frac{\partial \boldsymbol{u}}{\partial \nu} &= \boldsymbol{g}_1 \quad \text{on } \Gamma_1, \end{aligned}$$

where α and μ are given positive constants, $\boldsymbol{f}$, $\boldsymbol{g}_0$ and $\boldsymbol{g}_1$ are given functions, suitably smooth, $\partial\Omega = \Gamma_0 \cup \Gamma_1$ with $\Gamma_0 \cap \Gamma_1 = \emptyset$, and $\partial/\partial\nu$ denotes the outward normal derivative on the boundary. Derive the Newton iteration formula for solving this nonlinear problem.

Exercise 5.4.3 Explore sufficient conditions for the convergence of the Newton method (5.4.7).

Exercise 5.4.4 Explore sufficient conditions for the convergence of the Newton method (5.4.10).

Exercise 5.4.5 Explore sufficient conditions for the convergence of the modified Newton method (5.4.11)–(5.4.12).
Hint: Use the Banach fixed-point theorem to show the method is linearly convergent, provided u_0 is chosen sufficiently close to u^*.

5.5 Completely continuous vector fields

There are other means of asserting the existence of a solution to an equation. For example, if $f \in C[a,b]$, and if $f(a)\,f(b) < 0$, then the intermediate value theorem asserts the existence of a solution in $[a,b]$ to the equation $f(x) = 0$. We convert this to an existence theorem for fixed points as follows. Let $T : [a,b] \to [a,b]$ be continuous. Then $x = T(x)$ has a solution in $[a,b]$. This can be proved by reducing it to the earlier case, letting $f(x) \equiv x - T(x)$.

It is natural to try to extend this to multivariate functions f or T.

Theorem 5.5.1 (Brouwer's Fixed-Point Theorem) *Let $K \subset \mathbb{R}^d$ be bounded, closed, and convex. Let $T : K \to K$ be continuous. Then T has at least one fixed point in the set K.*

For a proof of Theorem 5.5.1, see [147, p. 94], [135, pp. 636–639], or [84, p. 232].

We would like to generalize this further, to operators on infinite dimensional Banach spaces. There are several ways of doing this, and we describe two approaches, both based on the assumption that T is a *continuous compact operator*. This is a concept we generalize to nonlinear operators T from Definition 2.8.1 for linear operators. We begin with an example to show that some additional hypotheses are needed, in addition to those assumed in Theorem 5.5.1.

Example 5.5.2 Assume V is a Hilbert space with an orthonormal basis $\{\varphi_j\}_{j\geq 1}$. Then for every $v \in V$, we can write

$$v = \sum_{j=1}^{\infty} \alpha_j \varphi_j, \quad \|v\|_V = \sqrt{\sum_{j=1}^{\infty} |\alpha_j|^2}.$$

Let K be the unit ball in V:

$$K = \{v \mid \|v\|_V \leq 1\}.$$

Introduce a parameter $k > 1$, and then choose a second parameter $t < 1$ satisfying

$$0 < t \leq \sqrt{k^2 - 1}.$$

Define $T : K \to K$ by

$$T(v) = t\,(1 - \|v\|_V)\,\varphi_1 + \sum_{j=1}^{\infty} \alpha_j \varphi_{j+1}, \quad v \in K. \tag{5.5.1}$$

This can be shown to be Lipschitz continuous on K, with

$$\|T(v) - T(w)\|_V \leq k\,\|v - w\|_V, \quad v, w \in K. \tag{5.5.2}$$

Moreover, the domain K is convex, closed, and bounded. However, T does not have a fixed point.

This example is a modification of one given in [133]. □

Definition 5.5.3 *Let $T : K \subset V \to W$, with V and W Banach spaces. We say T is* compact *if for every bounded set $B \subset K$, the set $T(B)$ has compact closure in W. If T is both compact and continuous, we call T a* completely continuous operator.

When T is a linear operator, T being compact implies T is bounded and hence continuous. This is not true in general when T is nonlinear; continuity of T must be assumed separately. Some authors include a requirement of continuity in their definition of T being compact, e.g. [33, p. 89]. With the above definition, we can state one generalization of Theorem 5.5.1. For proofs, see [33, p. 90], [135, p. 482], or [147, p. 124].

Theorem 5.5.4 (Schauder's fixed-point theorem) *Let V be a Banach space and let $K \subset V$ be bounded, closed, and convex. Assume $T : K \to K$ is a completely continuous operator. Then T has at least one fixed point in the set K.*

When dealing with equations involving differentiable nonlinear functions, say

$$v = T(v), \tag{5.5.3}$$

a common approach is to "linearize the problem". This generally means we replace the nonlinear function by a linear Taylor series approximation,

$$T(v) \approx T(v_0) + T'(v_0)\,(v - v_0) \tag{5.5.4}$$

for some suitably chosen point v_0. Then the equation (5.5.3) can be rewritten as

$$(I - T'(v_0))\,(v - v_0) \approx T(v_0) - v_0. \tag{5.5.5}$$

This linearization procedure is a commonly used approach for the convergence analysis of approximation methods for solving (5.5.3), and it is used to this end in Section 12.7. It is also the basis of Newton's method for (5.5.3).

The approximation (5.5.5) leads us to consider the properties of $T'(v_0)$ and motivates consideration of the following result. A proof is given in [148, p. 77]. As a consequence, we can apply the Fredholm alternative theorem to the operator $I - T'(v_0)$.

Proposition 5.5.5 *Let V be a Banach space and let $K \subset V$ be an open set. Assume $T : K \to V$ is a completely continuous operator which is differentiable at $v_0 \in K$. Then $T'(v_0)$ is a compact operator from V to V.*

5.5.1 The rotation of a completely continuous vector field

The concept of the rotation of a nonlinear mapping is a fairly deep and sophisticated consequence of topology, and a complete development of it is given in [147, Chap. 2]. We describe here the main properties of this "rotation".

Let $T : K \subset V \to V$, with V a Banach space, and assume T is completely continuous on K. We call the function

$$\Phi(v) \equiv v - T(v), \quad v \in K$$

the *completely continuous vector field generated by* T. Let B be a bounded, open subset of K and let S denote its boundary, and assume $\overline{B} \equiv B \cup S \subset K$. *Assume T has no fixed points on the boundary S.* Under the above assumptions, it is possible to define *the rotation of* T (or Φ) *over* S. This is an integer, denoted here by $\mathrm{Rot}(\Phi)$ with the following properties.

P1. If $\text{Rot}(\Phi) \neq 0$, then T has at least one fixed point within the set B. (See [147, p. 123].)

P2. Assume there is a function $X(v,t)$ defined for $v \in \overline{B}$ and $0 \leq t \leq 1$, and assume it has the following properties.

(a) $X(v,0) \equiv \Phi(v)$, $v \in \overline{B}$.

(b) $X(\cdot,t)$ is completely continuous on $\overline{B}$ for each $t \in [0,1]$.

(c) For every $v \in S$, $X(v,t)$ is uniformly continuous in t.

(d) $v - X(v,t) \neq 0$ for all $v \in S$ and for $0 \leq t \leq 1$.

Then $\text{Rot}(\Phi) = \text{Rot}(\Psi)$, where $\Psi(v) \equiv v - X(v,1)$. The mapping X is called a *homotopy*, and this property says *the rotation of a completely continuous vector field is invariant under homotopy.* (See [147, p. 108].)

P3. Let v_0 be an isolated fixed point of T in K. Then for all sufficiently small neighborhood of v_0, $\text{Rot}(\Phi)$ over that neighborhood is constant; it is called the *index of the fixed point* v_0. If all fixed points of T on B are isolated, then the number of such fixed points is finite; call them $v_1, \ldots, v_r$. Moreover, $\text{Rot}(\Phi)$ equals the sum of the indexes of the individual fixed points $v_1, \ldots, v_r$. (See [147, p. 109].)

P4. Let v_0 be a fixed point of T and suppose that T has a continuous Fréchet derivative $T'(v)$ for all v in some neighborhood of v_0. In addition, assume 1 is not an eigenvalue of $T'(v_0)$. Then the index of v_0 in nonzero. More precisely, it equals $(-1)^\beta$ with β equal to the number of positive real eigenvalues of $T'(v_0)$ which are greater than 1, counted according to their multiplicity. Also, the fixed point v_0 is isolated. (See [147, p. 136].)

P5. Let v_0 be an isolated fixed point of T in B. Then the index of v_0 is zero if and only if there exists some open neighborhood N of v_0 such that for every $\delta > 0$, there exists completely continuous T_δ defined on $\overline{N}$ to V with

$$\|T(v) - T_\delta(v)\|_V \leq \delta, \quad v \in \overline{N}$$

and with T_δ having no fixed points in $\overline{N}$. This says that isolated fixed points have index zero if and only if they are unstable with respect to completely continuous perturbations.

These ideas give a framework for the error analysis of numerical methods for solving some nonlinear integral equations and other problems. Such methods are examined in Subsection 12.7.2 of Chapter 12, and an example is given in [12].

Exercise 5.5.1 Returning to Example 5.5.2, show that $T : B \to B$, the inequality (5.5.2) holds, and T does not have a fixed point in B.

5.6 Conjugate gradient method for operator equations

The *conjugate gradient method* is an iteration method which was originally devised for solving finite linear systems which were symmetric and positive definite. The method has since been generalized in a number of directions, and in this section, we consider its generalization to operator equations

$$Au = f. \tag{5.6.1}$$

In Section 9.4, we consider the variational formulation of the conjugate gradient method. Here, we assume A is a bounded, positive definite, self-adjoint linear operator on a Hilbert space V. With these assumptions, we can apply Theorem 5.1.4 to conclude that (5.6.1) has a unique solution $u^* = A^{-1}f$ and the inverse operator A^{-1} is continuous from V to V. For simplicity in this section, we assume V is a real Hilbert space; and we further assume that V is *separable*, implying that it has a countable orthogonal basis.

The conjugate gradient method for solving $Au = f$ in V is defined as follows. Let u_0 be an initial guess for the solution u^*. Define $r_0 = f - Au_0$ and $s_0 = r_0$. For $k \geq 0$, define

$$\begin{aligned} u_{k+1} &= u_k + \alpha_k s_k, & \alpha_k &= \frac{\|r_k\|^2}{(As_k, s_k)}, \\ r_{k+1} &= f - Au_{k+1}, & & \\ s_{k+1} &= r_{k+1} + \beta_k s_k, & \beta_k &= \frac{\|r_{k+1}\|^2}{\|r_k\|^2}. \end{aligned} \tag{5.6.2}$$

The norm and inner product are those of V. There are several other equivalent formulations of the method. An introduction to the conjugate gradient method for finite-dimensional systems, together with some other equivalent ways of writing it, is given in [15, Section 8.9] and [95, Section 10.2].

The following theorem is taken from [186, p. 159]; and we omit the proof, as it calls for a more detailed investigation of the method than we wish to consider here. The proof also follows quite closely the proof for the finite-dimensional case, which is well-known in the literature (e.g. see [156, p. 250]). In stating the theorem, we also use the following alternative inner product and norm:

$$(v, u)_A = (Av, u), \quad \|v\|_A = \sqrt{(v, v)_A}.$$

Theorem 5.6.1 *Let A be a bounded, self-adjoint, linear operator satisfying*

$$\sqrt{m}\,\|v\| \le \|v\|_A \le \sqrt{M}\,\|v\|, \quad v \in V, \tag{5.6.3}$$

with $m, M > 0$ (which implies that $\|\cdot\|_A$ and $\|\cdot\|$ are equivalent norms). Then the sequence $\{u_k\}$ from (5.6.2) converges to u^*, and

$$\|u^* - u_{k+1}\|_A \le \frac{M-m}{M+m}\,\|u^* - u_k\|_A\,, \quad k \ge 0. \tag{5.6.4}$$

This shows $u_k \to u^*$ linearly.

Patterson [186, p. 163] also derives the improved result

$$\|u^* - u_k\|_A \le 2\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^k \|u^* - u_0\|_A\,, \quad k \ge 0. \tag{5.6.5}$$

It follows that this is a more rapid rate of convergence by showing

$$\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}} \le \frac{M-m}{M+m}, \tag{5.6.6}$$

which we leave to the reader.

In the case that $A = I - K$, with K a compact operator, more can be said about the rate of convergence of u_k to u^*. For the remainder of this section, we assume K is a compact, self-adjoint operator on V to V.

The discussion of the convergence requires results on the eigenvalues of K. From Theorem 2.8.15 in Subsection 2.8.5, the eigenvalues of the self-adjoint compact operator K are real and the associated eigenvectors can be so chosen as to form an orthonormal basis for V:

$$K\phi_j = \lambda_j \phi_j, \quad j = 1, 2, \ldots$$

with $(\phi_i, \phi_j) = \delta_{i,j}$. Without any loss of generality, let the eigenvalues be ordered as follows:

$$|\lambda_1| \ge |\lambda_2| \ge |\lambda_3| \ge \cdots \ge 0. \tag{5.6.7}$$

We permit the number of nonzero eigenvalues to be finite or infinite. From Theorem 2.8.12 of Subsection 2.8.5,

$$\lim_{j\to\infty} \lambda_j = 0.$$

The eigenvalues of A are $\{1 - \lambda_j\}_{j\ge1}$ with $\{\phi_j\}_{j\ge1}$ as the corresponding orthogonal eigenvectors. The self-adjoint operator $A = I - K$ is positive definite if and only if

$$\delta \equiv \inf_{j\ge1}(1-\lambda_j) = 1 - \sup_{j\ge1}\lambda_j > 0, \tag{5.6.8}$$

or equivalently, $\lambda_j < 1$ for all $j \geq 1$. For later use, also introduce

$$\Delta \equiv \sup_{j \geq 1} (1 - \lambda_j) .$$

We note that

$$\|A\| = \Delta, \quad \left\|A^{-1}\right\| = \frac{1}{\delta}. \tag{5.6.9}$$

With respect to (5.6.3), $M = \Delta$ and $m = \delta$. The result (5.6.4) becomes

$$\|u^* - u_{k+1}\|_A \leq \frac{\Delta - \delta}{\Delta + \delta} \|u^* - u_k\|_A , \quad k \geq 0. \tag{5.6.10}$$

It is possible to improve on this geometric rate of convergence when dealing with equations

$$Au \equiv (I - K) u = f$$

to show a "superlinear" rate of convergence. The following result is due to Winther [232].

Theorem 5.6.2 *Let K be a self-adjoint compact operator on the Hilbert space V. Assume $A = I - K$ is a self-adjoint positive definite operator (with the notation used above). Let $\{u_k\}$ be generated by the conjugate gradient iteration* (5.6.2)*. Then $u_k \to u^*$ superlinearly:*

$$\|u^* - u_k\| \leq (c_k)^k \|u^* - u_0\| , \quad k \geq 0 \tag{5.6.11}$$

with $\lim_{k \to \infty} c_k = 0$.

Proof. It is a standard result (see [156, p. 246]) that (5.6.2) implies that

$$u_k = u_0 + \widetilde{P}_{k-1}(A) r_0 \tag{5.6.12}$$

with $\widetilde{P}_{k-1}(\lambda)$ a polynomial of degree $\leq k - 1$. Letting $A = I - K$, this can be rewritten in the equivalent form

$$u_k = u_0 + \widehat{P}_{k-1}(K) r_0$$

for some other polynomial $\widehat{P}(\lambda)$ of degree $\leq k - 1$. The conjugate gradient iterates satisfy an optimality property: If $\{y_k\}$ is another sequence of iterates, generated by another sequence of the form

$$y_k = u_0 + P_{k-1}(K) r_0, \quad k \geq 1, \tag{5.6.13}$$

for some sequence of polynomials $\{P_{k-1} \mid \deg(P_{k-1}) \leq k - 1,\ k \geq 1\}$, then

$$\|u^* - u_k\|_A \leq \|u^* - y_k\|_A , \quad k \geq 0. \tag{5.6.14}$$

For a proof, see Luenberger [156, p. 247].

Introduce

$$Q_k(\lambda) = \prod_{j=1}^{k} \frac{\lambda - \lambda_j}{1 - \lambda_j}, \tag{5.6.15}$$

and note that $Q_k(1) = 1$. Define P_{k-1} implicitly by

$$Q_k(\lambda) = 1 - (1 - \lambda)P_{k-1}(\lambda),$$

and note that $\deg(P_{k-1}) = k-1$. Let $\{y_k\}$ be defined using (5.6.13). Define $\widetilde{e}_k = u^* - y_k$ and

$$\widetilde{r}_k = b - Ay_k = A\widetilde{e}_k.$$

We first bound $\widetilde{r}_k$, and then

$$\|\widetilde{e}_k\| \le \left\|A^{-1}\right\| \|\widetilde{r}_k\| = \frac{1}{\delta} \|\widetilde{r}_k\|.$$

Moreover,

$$\|u^* - u_k\|_A \le \|u^* - y_k\|_A \le \sqrt{\Delta}\, \|u^* - y_k\| \le \frac{\sqrt{\Delta}}{\delta} \|\widetilde{r}_k\|;$$

hence,

$$\|u^* - u_k\| \le \frac{1}{\sqrt{\delta}} \|u^* - u_k\|_A \le \frac{1}{\delta}\sqrt{\frac{\Delta}{\delta}} \|\widetilde{r}_k\|. \tag{5.6.16}$$

From (5.6.13),

$$\widetilde{r}_k = b - A\left[y_0 + P_{k-1}(K)r_0\right] = \left[I - AP_{k-1}(K)\right] r_0 = Q_k(A)r_0. \tag{5.6.17}$$

Expand r_0 using the eigenfunction basis $\{\phi_1, \phi_2, \dots\}$:

$$r_0 = \sum_{j=1}^{\infty} (r_0, \phi_j)\phi_j.$$

Note that

$$Q_k(A)\phi_j = Q_k(\lambda_j)\phi_j, \quad j \ge 1,$$

and thus $Q_k(A)\phi_j = 0$ for $j = 1, \dots, k$. Then (5.6.17) implies

$$\widetilde{r}_k = \sum_{j=1}^{\infty} (r_0, \phi_j) Q_k(A)\phi_j = \sum_{j=k+1}^{\infty} (r_0, \phi_j) Q_k(\lambda_j)\phi_j$$

and

$$\|\widetilde{r}_k\| \le \alpha_k \sqrt{\sum_{j=k+1}^{\infty} (r_0, \phi_j)^2} \le \alpha_k \|r_0\| \tag{5.6.18}$$

with

$$\alpha_k = \sup_{j \geq k+1} |Q_k(\lambda_j)| .$$

Examining $Q_k(\lambda_j)$ and using (5.6.7), we have

$$\alpha_k \leq \prod_{j=1}^{k} \frac{|\lambda_{k+1}| + |\lambda_j|}{1 - \lambda_j} \leq \prod_{j=1}^{k} \frac{2\,|\lambda_j|}{1 - \lambda_j}. \tag{5.6.19}$$

Recall the well-known inequality

$$\left(\prod_{j=1}^{k} b_j \right)^{1/k} \leq \frac{1}{k} \sum_{j=1}^{k} b_j$$

which relates the arithmetic and geometric means of k positive numbers $b_1, \ldots, b_k$. Applying this inequality to the right hand side of (5.6.19), we have

$$\alpha_k \leq \left(\frac{2}{k} \sum_{j=1}^{k} \frac{|\lambda_j|}{1 - \lambda_j} \right)^k .$$

Since $\lim_{j \to \infty} \lambda_j = 0$, it is a straightforward argument to show that

$$\lim_{k \to \infty} \frac{2}{k} \sum_{j=1}^{k} \frac{|\lambda_j|}{1 - \lambda_j} = 0. \tag{5.6.20}$$

We leave the proof to the reader.

Returning to (5.6.18) and (5.6.16), we have

$$\|u^* - u_k\| \leq \frac{1}{\delta} \sqrt{\frac{\Delta}{\delta}}\, \|\widetilde{r}_k\| \leq \frac{1}{\delta} \sqrt{\frac{\Delta}{\delta}}\, \alpha_k \|r_0\| \leq \left(\frac{\Delta}{\delta} \right)^{3/2} \alpha_k \|u^* - u_0\|.$$

To obtain (5.6.11), define

$$c_k = \left(\frac{\Delta}{\delta} \right)^{3/(2k)} \left(\frac{2}{k} \sum_{j=1}^{k} \frac{|\lambda_j|}{1 - \lambda_j} \right). \tag{5.6.21}$$

By (5.6.20), $c_k \to 0$ as $k \to \infty$. □

It is of interest to know how rapidly c_k converges to zero. For this we restrict ourselves to compact integral operators. For simplicity, we consider only the single variable case:

$$Kv(x) = \int_a^b k(x,y) v(y)\, dy, \quad x \in [a,b],\ v \in L^2(a,b), \tag{5.6.22}$$

and $V = L^2(a,b)$. From (5.6.21), the speed of convergence of $c_k \to 0$ is essentially the same as that of

$$\tau_k \equiv \frac{1}{k}\sum_{j=1}^{k}\frac{|\lambda_j|}{1-\lambda_j}. \tag{5.6.23}$$

In turn, the convergence of τ_k depends on the rate at which the eigenvalues λ_j converge to zero. We give two results from Flores [82] in the following theorem. In all cases, we also assume the operator $A = I - K$ is positive definite, which is equivalent to the assumption (5.6.8).

Theorem 5.6.3 (a) *Assume the integral operator K of* (5.6.22) *is a self-adjoint Hilbert-Schmidt integral operator, i.e.*

$$\|K\|_{HS}^2 \equiv \int_a^b \int_a^b |k(t,s)|^2 ds\, dt < \infty.$$

Then

$$\frac{1}{\ell}\cdot\frac{|\lambda_1|}{1-\lambda_1} \le \tau_\ell \le \frac{1}{\sqrt{\ell}}\|K\|_{HS}\left\|(I-K)^{-1}\right\|. \tag{5.6.24}$$

(b) *Assume $k(t,s)$ is a symmetric kernel with continuous partial derivatives of order up to p for some $p \ge 1$. Then there is a constant $M \equiv M(p)$ such that*

$$\tau_\ell \le \frac{M}{\ell}\zeta(p+1/2)\left\|(I-K)^{-1}\right\|, \quad \ell \ge 1 \tag{5.6.25}$$

with $\zeta(z)$ the Riemann zeta function:

$$\zeta(z) = \sum_{m=1}^{\infty}\frac{1}{m^z}, \quad z > 1.$$

Proof. (a) It can be proven from Theorem 2.8.15 of Section 2.8.5 that

$$\sum_{j=1}^{\infty}\lambda_j^2 = \|K\|_{HS}^2.$$

From this, the eigenvalues λ_j can be shown to converge to zero with a certain speed. Namely,

$$j\lambda_j^2 \le \sum_{i=1}^{j}\lambda_i^2 \le \|K\|_{HS}^2,$$

and so

$$\lambda_j \le \frac{1}{\sqrt{j}}\|K\|_{HS}, \quad j \ge 1.$$

This leads to

$$\tau_\ell = \frac{1}{\ell}\sum_{j=1}^{\ell}\frac{|\lambda_j|}{1-\lambda_j} \le \frac{1}{\delta}\frac{\|K\|_{HS}}{\ell}\sum_{j=1}^{\ell}\frac{1}{\sqrt{j}} \le \frac{1}{\sqrt{\ell}}\frac{\|K\|_{HS}}{\delta}.$$

Recalling that $\delta^{-1} = \left\|A^{-1}\right\|$ proves the upper bound in (5.6.24). The lower bound in (5.6.24) is immediate from the definition of τ_ℓ.
(b) From Fenyö and Stolle [80, Sec. 8.9], the eigenvalues $\{\lambda_j\}$ satisfy

$$\lim_{j\to\infty} j^{p+1/2}\lambda_j = 0.$$

Let

$$\beta \equiv \sup_j j^{p+1/2}|\lambda_j|$$

so that

$$|\lambda_j| \le \frac{\beta}{j^{p+1/2}}, \quad j \ge 1. \tag{5.6.26}$$

With this bound on the eigenvalues,

$$\tau_\ell \le \frac{\beta}{\ell\delta}\sum_{j=1}^{\ell}\frac{1}{j^{p+1/2}} \le \frac{\beta}{\ell\delta}\zeta(p+1/2).$$

This completes the proof of (5.6.25). □

We see that the speed of convergence of $\{\tau_\ell\}$ (or equivalently, $\{c_\ell\}$) is no better that $\mathcal{O}(\ell^{-1})$, regardless of the differentiability of the kernel function k. Moreover, for most cases of practical interest, it is no worse than $\mathcal{O}(\ell^{-1/2})$. The result (5.6.11) was only a bound for the speed of convergence of the conjugate gradient method, although we expect the convergence speed is no better than this. For additional discussion of the convergence of $\{\tau_\ell\}$, see [82].

The result (5.6.12) says that the vectors $u_k - u_0$ belong to the *Krylov subspace*

$$\mathcal{K}(A) \equiv \left\{r_0, Ar_0, A^2r_0, \ldots, A^{k-1}r_0\right\};$$

and in fact, u_k is an optimal choice in the following sense:

$$\|u^* - u_k\|_A = \min_{y\in u_0+\mathcal{K}(A)}\|u^* - y\|_A.$$

Other iteration methods have been based on choosing particular elements from $\mathcal{K}(A)$ using a different sense of optimality. For a general discussion of such generalizations to nonsymmetric finite linear systems, see [87]. The conjugate gradient method has also been extended to the solution of nonlinear systems and nonlinear optimization problems.

Exercise 5.6.1 Prove the inequality (5.6.6).

Exercise 5.6.2 Derive the relation (5.6.12).

Exercise 5.6.3 Prove the result in (5.6.20).

Suggestion for Further Reading.

Many books cover the convergence issue of iterative methods for finite linear systems; see, e.g., Axelsson [28], Demmel [67], Golub and Van Loan [95], Stewart [214], and Trefethen and Bau [224]. For finite-dimensional nonlinear systems, see the comprehensive work of Ortega and Rheinboldt [184]. The recent book by Deuflhard [68] focuses on affine invariance properties of the Newton method and its variants for algorithms and their convergence analysis, and this approach permits the construction of fully adaptive algorithms. For optimization problems, two comprehensive references are Luenberger [156] and Nocedal and Wright [181]. The conjugate gradient method was first published in [122] for solving symmetric positive-definite linear systems. The method and its extensions for more general problems, mainly finite dimensional ones, are discussed in several books, e.g., Golub and Van Loan [95], Hestenes [121], Luenberger [156], Nocedal and Wright [181].

Portions of this chapter follow Zeidler [244]. A more theoretical look at iteration methods for solving linear equations is given in Nevanlinna [178]. The iterative solution of linear integral equations of the second kind is treated in several books, including Atkinson [18, Chap. 6] and Kress [149]. There are many other tools for the analysis of nonlinear problems, and we refer the reader to Berger [33], Franklin [84], Kantorovich and Akilov [135, Chaps. 16–18], Krasnoselskii [147], Krasnoselskii and Zabreyko [148], and Zeidler [244].

6
Finite Difference Method

The finite difference method is a universally applicable numerical method for the solution of differential equations. In this chapter, for a sample parabolic partial differential equation, we introduce some difference schemes and analyze their convergence. We present the well-known Lax equivalence theorem and related theoretical results, and apply them to the convergence analysis of difference schemes.

The finite difference method can be difficult to analyze, in part because it is quite general in its applicability. Much of the existing stability and convergence analysis is restricted to special cases, particularly to linear differential equations with constant coefficients. These results are then used to predict the behavior of difference methods for more complicated equations.

6.1 Finite difference approximations

The basic idea of the finite difference method is to approximate differential quotients by appropriate difference quotients, thus reducing a differential equation to an algebraic system. There are a variety of ways to do the approximation.

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_6,
© Springer Science + Business Media, LLC 2009

Suppose f is a differentiable real-valued function on $\mathbb{R}$. Let $x \in \mathbb{R}$ and $h > 0$. Then we have the following three popular difference approximations:

$$f'(x) \approx \frac{f(x+h) - f(x)}{h} \tag{6.1.1}$$

$$\frac{f(x) - f(x-h)}{h} \tag{6.1.2}$$

$$\frac{f(x+h) - f(x-h)}{2h}. \tag{6.1.3}$$

These differences are called a *forward difference*, a *backward difference* and a *centered difference*, respectively. Supposing f has a second derivative, it is easy to verify that the approximation errors for the forward and backward differences are both $\mathcal{O}(h)$. If the third derivative of f exists, then the approximation error for the centered difference is $\mathcal{O}(h^2)$. We see that if the function is smooth, the centered difference is a more accurate approximation to the derivative.

The second derivative of the function f is usually approximated by a second-order centered difference:

$$f''(x) \approx \frac{f(x+h) - 2\,f(x) + f(x-h)}{h^2}. \tag{6.1.4}$$

It can be verified that when f has a fourth derivative, the approximation error is $O(h^2)$.

Now let us use these difference formulas to formulate some difference schemes for a sample initial-boundary value problem for a heat equation.

Example 6.1.1 Let us consider the problem

$$u_t = \nu\, u_{xx} + f(x,t) \quad \text{in } (0,\pi) \times (0,T), \tag{6.1.5}$$

$$u(0,t) = u(\pi,t) = 0, \quad 0 \le t \le T, \tag{6.1.6}$$

$$u(x,0) = u_0(x), \quad 0 \le x \le \pi. \tag{6.1.7}$$

The differential equation (6.1.5) can be used to model a variety of physical processes such as heat conduction (see e.g. [192]). Here $\nu > 0$ is a given constant, f and u_0 are given continuous functions. To develop a finite difference method, we need to introduce grid points. Let N_x and N_t be positive integers, $h_x = \pi/N_x$, $h_t = T/N_t$ and define the partition points

$$x_j = j\,h_x, \quad j = 0, 1, \ldots, N_x,$$
$$t_m = m\,h_t, \quad m = 0, 1, \ldots, N_t.$$

A point of the form (x_j, t_m) is called a *grid point* and we are interested in computing approximate solution values at the grid points. We use the notation v_j^m for an approximation to $u_j^m \equiv u(x_j, t_m)$ computed from a finite difference scheme. Write $f_j^m = f(x_j, t_m)$ and

$$r = \nu\, h_t / h_x^2.$$

Then we can bring in several schemes.

The first scheme is

$$\frac{v_j^{m+1} - v_j^m}{h_t} = \nu \frac{v_{j+1}^m - 2v_j^m + v_{j-1}^m}{h_x^2} + f_j^m,$$
$$1 \le j \le N_x - 1,\ 0 \le m \le N_t - 1, \tag{6.1.8}$$
$$v_0^m = v_{N_x}^m = 0, \qquad 0 \le m \le N_t, \tag{6.1.9}$$
$$v_j^0 = u_0(x_j), \qquad 0 \le j \le N_x. \tag{6.1.10}$$

This scheme is obtained by discretizing the differential equation (6.1.5) at $x = x_j$ and $t = t_m$, replacing the time derivative with a forward difference and the second spatial derivative with a second-order centered difference. Hence it is called a forward-time centered-space scheme. The difference equation (6.1.8) can be written as

$$v_j^{m+1} = (1 - 2r)\, v_j^m + r\,(v_{j+1}^m + v_{j-1}^m) + h_t f_j^m,$$
$$1 \le j \le N_x - 1,\ 0 \le m \le N_t - 1. \tag{6.1.11}$$

Thus once the solution at the time level $t = t_m$ is computed, the solution at the next time level $t = t_{m+1}$ can be found explicitly. The forward scheme (6.1.8)–(6.1.10) is an *explicit method*.

Alternatively, we may replace the time derivative with a backward difference and still use the second order centered difference for the second spatial derivative. The resulting scheme is a backward-time centered-space scheme:

$$\frac{v_j^m - v_j^{m-1}}{h_t} = \nu \frac{v_{j+1}^m - 2v_j^m + v_{j-1}^m}{h_x^2} + f_j^m,$$
$$1 \le j \le N_x - 1,\ 1 \le m \le N_t, \tag{6.1.12}$$
$$v_0^m = v_{N_x}^m = 0, \qquad 0 \le m \le N_t, \tag{6.1.13}$$
$$v_j^0 = u_0(x_j), \qquad 0 \le j \le N_x. \tag{6.1.14}$$

The difference equation (6.1.12) can be written as

$$(1 + 2r)\, v_j^m - r\,(v_{j+1}^m + v_{j-1}^m) = v_j^{m-1} + h_t f_j^m,$$
$$1 \le j \le N_x - 1,\ 1 \le m \le N_t, \tag{6.1.15}$$

which is supplemented by the boundary condition from (6.1.13). Thus in order to find the solution at the time level $t = t_m$ from the solution at $t = t_{m-1}$, we need to solve a tridiagonal linear system of order $N_x - 1$. The backward scheme (6.1.12)–(6.1.14) is an *implicit method*.

In the above two methods, we approximate the differential equation at $x = x_j$ and $t = t_m$. We can also consider the differential equation at $x = x_j$ and $t = t_{m-1/2}$, approximating the time derivative by a centered difference:

$$u_t(x_j, t_{m-1/2}) \approx \frac{u(x_j, t_m) - u(x_j, t_{m-1})}{h_t}.$$

Further, approximate the second spatial derivative by the second order centered difference:

$$u_{xx}(x_j, t_{m-1/2}) \approx \frac{u(x_{j+1}, t_{m-1/2}) - 2\,u(x_j, t_{m-1/2}) + u(x_{j-1}, t_{m-1/2})}{h_t^2},$$

and then approximate the half time values by averages:

$$u(x_j, t_{m-1/2}) \approx [u(x_j, t_m) + u(x_j, t_{m-1})]\,/2,$$

etc. As a result we arrive at the *Crank-Nicolson scheme*:

$$\frac{v_j^m - v_j^{m-1}}{h_t} = \nu\,\frac{(v_{j+1}^m - 2v_j^m + v_{j-1}^m) + (v_{j+1}^{m-1} - 2v_j^{m-1} + v_{j-1}^{m-1})}{2\,h_x^2} + f_j^{m-1/2}, \qquad 1 \le j \le N_x - 1,\ 1 \le m \le N_t, \tag{6.1.16}$$

$$v_0^m = v_{N_x}^m = 0, \qquad 0 \le m \le N_t, \tag{6.1.17}$$

$$v_j^0 = u_0(x_j), \qquad 0 \le j \le N_x. \tag{6.1.18}$$

Here $f_j^{m-1/2} = f(x_j, t_{m-1/2})$, which can be replaced by $(f_j^m + f_j^{m-1})/2$. The difference equation (6.1.16) can be rewritten as

$$(1+r)\,v_j^m - \frac{r}{2}\,(v_{j+1}^m + v_{j-1}^m) = (1-r)\,v_j^{m-1} + \frac{r}{2}\,(u_{j+1}^{m-1} + u_{j-1}^{m-1}) + h_t f_j^{m-1/2}. \tag{6.1.19}$$

We see that the Crank-Nicolson scheme is also an implicit method and at each time step we need to solve a tridiagonal linear system of order $N_x - 1$.

The three schemes derived above all seem reasonable approximations for the initial-boundary value problem (6.1.5)–(6.1.7). Let us do some numerical experiments to see if these schemes indeed produce useful results. Let us use the forward scheme (6.1.8)–(6.1.10) and the backward scheme (6.1.12)–(6.1.14) to solve the problem (6.1.5)–(6.1.7) with $\nu = 1$, $f(x,t) = 0$ and $u_0(x) = \sin x$. It can be verified that the exact solution is $u(x,t) = e^{-t}\sin x$. We consider numerical solution errors at $t = 1$. Results from the Crank-Nicolson scheme are qualitatively similar to those from the backward scheme but magnitudes of the error are smaller, and are thus omitted.

Figure 6.1 shows solution errors of the forward scheme corresponding to several combinations of the values N_x and N_t (or equivalently, h_x and h_t). Convergence is observed only when N_x is substantially smaller than N_t (i.e. when h_t is substantially smaller than h_x). In the next two sections, we explain this phenomenon theoretically.

Figure 6.2 demonstrates solution errors of the backward scheme corresponding to the same values of N_x and N_t. We observe a good convergence pattern. The maximum solution error decreases as N_x and N_t increase. In

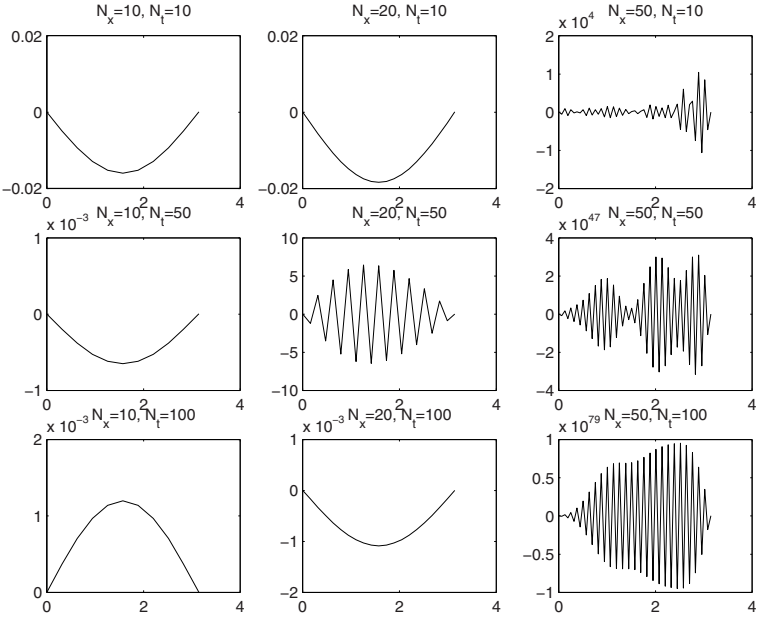

FIGURE 6.1. The forward scheme: errors at $t = 1$

Section 6.3, we prove that the maximum error at $t = 1$ is bounded by a constant times $(h_x^2 + h_t)$. This result explains the phenomenon in Figure 6.2 that the error seems to decrease more rapidly with a decreasing h_t than h_x. □

Naturally, a difference scheme is useful only if the scheme is convergent, i.e., if it can provide numerical solutions which approximate the exact solution. A necessary requirement for convergence is consistency of the scheme, that is, the difference scheme must be close to the differential equation in some sense. However, consistency alone does not guarantee the convergence, as we see from the numerical examples above. From the view-point of theoretical analysis, at each time level, some error is brought in, representing the discrepancy between the difference scheme and the differential equation. From the view-point of computer implementation, numerical values and numerical computations are subject to roundoff errors. Thus it is important to be able to control the propagation of errors. The ability to control the propagation of errors is termed *stability* of the scheme. We expect to have convergence for consistent, stable schemes. The well-known Lax theory for finite difference methods goes beyond this. The theory states that with properly defined notions of consistency, stability and convergence

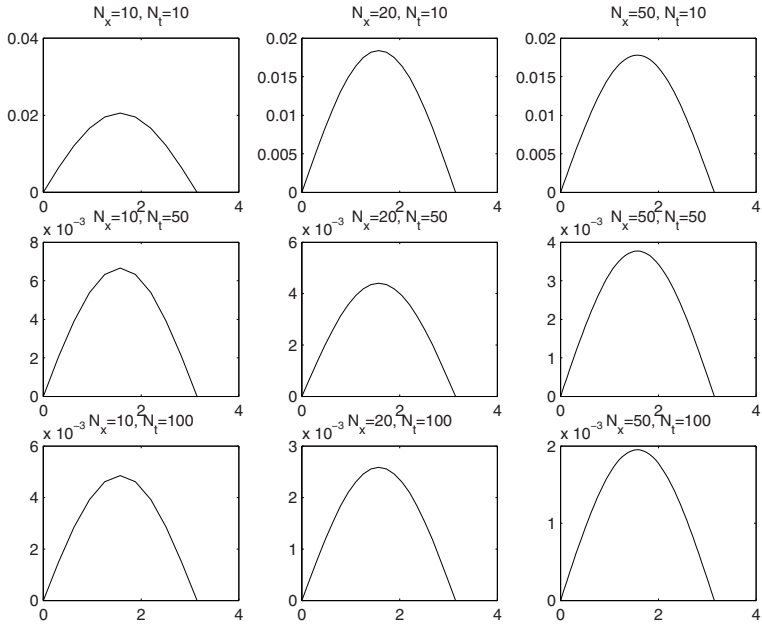

FIGURE 6.2. The backward scheme: errors at $t = 1$

for a well-posed partial differential equation problem, a consistent scheme is convergent if and only if it is stable. In the next section, we present one version of the Lax equivalence theory on the convergence of difference schemes. In the third section, we present and illustrate a variant of the Lax equivalence theory that is usually more easily applicable to yield convergence and convergence order results of difference schemes.

Exercise 6.1.1 One approach for deriving difference formulas to approximate derivatives is the *method of undetermined coefficients*. Suppose f is a smooth function on $\mathbb{R}$. Let $h > 0$. Determine coefficients a, b, and c so that

$$a\, f(x+h) + b\, f(x) + c\, f(x-h)$$

is an approximation of $f'(x)$ with an order as high as possible; i.e., choose a, b, and c such that

$$\left| a\, f(x+h) + b\, f(x) + c\, f(x-h) - f'(x) \right| \leq \mathcal{O}(h^p)$$

with a largest possible exponent p.

Exercise 6.1.2 Do the same problem as Exercise 6.1.1 with $f'(x)$ replaced by $f''(x)$.

Exercise 6.1.3 Is it possible to use

$$a\,f(x+h)+b\,f(x)+c\,f(x-h)$$

with suitably chosen coefficients to approximate $f'''(x)$? How many function values are needed to approximate $f'''(x)$?

Exercise 6.1.4 For the initial value problem of the one-way wave equation

$$u_t + a\,u_x = f \qquad \text{in } \mathbb{R}\times\mathbb{R}_+, \tag{6.1.20}$$

$$u(\cdot,0) = u_0(\cdot) \qquad \text{in } \mathbb{R}, \tag{6.1.21}$$

where $a\in\mathbb{R}$ is a constant, derive some difference schemes based on various combinations of difference approximations of the time derivative and spatial derivative.

Exercise 6.1.5 The idea of the Lax-Wendroff scheme for solving the initial value problem of Exercise 6.1.4 is the following. Start with the Taylor expansion

$$u(x_j,t_{m+1}) \approx u(x_j,t_m) + h_t u_t(x_j,t_m) + \frac{h_t^2}{2}u_{tt}(x_j,t_m). \tag{6.1.22}$$

From the differential equation, we have

$$u_t = -a\,u_x + f$$

and

$$u_{tt} = a^2 u_{xx} - a\,f_x + f_t.$$

Use these relations to replace the time derivatives in the right side of (6.1.22). Then replace the first and the second spatial derivatives by central differences. Finally replace f_x by a central difference and f_t by a forward difference.

Follow the above instructions to derive the *Lax-Wendroff scheme* for solving (6.1.20)–(6.1.21).

Exercise 6.1.6 Give forward, backward and Crank-Nicolson schemes for the initial-boundary value problem

$$\begin{aligned}
u_t &= \nu_1 u_{xx} + \nu_2 u_{yy} + f(x,y,t), \quad 0<x<a,\ 0<y<b,\ 0<t<T,\\
u(0,y,t) &= g_l(y,t),\ u(a,y,t) = g_r(y,t), \quad 0\le y\le b,\ 0\le t\le T,\\
u(x,0,t) &= g_b(x,t),\ u(x,b,t) = g_t(x,t), \quad 0\le x\le a,\ 0\le t\le T,\\
u(x,y,0) &= u_0(x,y), \quad 0\le x\le a,\ 0\le y\le b.
\end{aligned}$$

Here for the given data, ν_1,ν_2,a,b,T are positive numbers, and f,g_l,g_r,g_b,g_t,u_0 are continuous functions of their arguments.

Exercise 6.1.7 Consider the initial-boundary value problem of a hyperbolic equation

$$\begin{aligned}
u_{tt} &= \nu\,u_{xx} + f(x,t) \quad \text{in } (0,a)\times(0,T),\\
u(0,t) &= g_1(t),\ u(a,t) = g_2(t), \quad 0\le t\le T,\\
u(x,0) &= u_0(x),\ u_t(x,0) = v_0(x), \quad 0\le x\le a.
\end{aligned}$$

Here for the given data, ν,a,T are positive numbers, and f,g_1,g_2,u_0,v_0 are continuous functions of their arguments. Discuss how to form a reasonable finite difference scheme for solving the problem.

6.2 Lax equivalence theorem

In this section, we follow [107] to present one version of the Lax equivalence theorem for analyzing difference methods in solving initial value or initial-boundary value problems. The rigorous theory is developed in an abstract setting. To help understand the theory, we use the sample problem (6.1.5)–(6.1.7) with $f(x,t) = 0$ to illustrate the notation, assumptions, definitions and the equivalence result.

We first introduce an abstract framework. Let V be a Banach space, $V_0 \subset V$ a dense subspace of V. Let $L : V_0 \subset V \to V$ be a linear operator. The operator L is usually unbounded and can be thought of as a differential operator. Consider the initial value problem

$$\begin{cases} \dfrac{du(t)}{dt} = Lu(t), \quad 0 \le t \le T, \\ u(0) = u_0. \end{cases} \tag{6.2.1}$$

This problem also represents an initial-boundary value problem with homogeneous boundary value conditions when they are included in definitions of the space V and the operator L. The next definition gives the meaning of a solution of the problem (6.2.1).

Definition 6.2.1 *A function $u : [0,T] \to V$ is a solution of the initial value problem* (6.2.1) *if for any $t \in [0,T]$, $u(t) \in V_0$,*

$$\lim_{\Delta t \to 0} \left\| \frac{1}{\Delta t} [u(t+\Delta t) - u(t)] - Lu(t) \right\| = 0, \tag{6.2.2}$$

and $u(0) = u_0$.

In the above definition, the limit in (6.2.2) is understood to be the right limit at $t = 0$ and the left limit at $t = T$.

Definition 6.2.2 *The initial value problem* (6.2.1) *is* well-posed *if for any $u_0 \in V_0$, there is a unique solution $u = u(t)$ and the solution depends continuously on the initial value: There exists a constant $c_0 > 0$ such that if $u(t)$ and $\overline{u}(t)$ are the solutions for the initial values $u_0, \overline{u}_0 \in V_0$, then*

$$\sup_{0 \le t \le T} \|u(t) - \overline{u}(t)\|_V \le c_0 \|u_0 - \overline{u}_0\|_V. \tag{6.2.3}$$

From now on, we assume the initial value problem (6.2.1) is well-posed. We denote the solution as

$$u(t) = S(t)\, u_0, \quad u_0 \in V_0.$$

Using the linearity of the operator L, it is easy to see that the solution operator $S(t)$ is linear. From the continuous dependence property (6.2.3),

we have

$$\sup_{0\le t\le T} \|S(t)\,(u_0-\overline{u}_0)\|_V \le c_0\|u_0-\overline{u}_0\|_V,$$
$$\sup_{0\le t\le T} \|S(t)u_0\|_V \le c_0\|u_0\|_V \quad \forall\, u_0\in V_0.$$

By Theorem 2.4.1, the operator $S(t) : V_0 \subset V \to V$ can be uniquely extended to a linear continuous operator $S(t) : V \to V$ with

$$\sup_{0\le t\le T} \|S(t)\|_V \le c_0.$$

Definition 6.2.3 *For $u_0 \in V\backslash V_0$, we call $u(t) = S(t)u_0$ the generalized solution of the initial value problem* (6.2.1).

Example 6.2.4 We use the following problem and its finite difference approximations to illustrate the use of the abstract framework of the section:

$$\begin{cases} u_t = \nu\, u_{xx} & \text{in } (0,\pi)\times(0,T), \\ u(0,t) = u(\pi,t) = 0 & 0\le t\le T, \\ u(x,0) = u_0(x) & 0\le x\le \pi. \end{cases} \tag{6.2.4}$$

We take $V = C_0[0,\pi] = \{v \in C[0,\pi] \mid v(0) = v(\pi) = 0\}$ with the norm $\|\cdot\|_{C[0,\pi]}$. We choose

$$V_0 = \Big\{ v \;\Big|\; v(x) = \sum_{j=1}^{n} a_j \sin(jx),\ a_j \in \mathbb{R}\,,\ n = 1,2,\dots \Big\}. \tag{6.2.5}$$

The verification that V_0 is dense in V is left as an exercise.

If $u_0 \in V_0$, then for some integer $n \ge 1$ and $b_1,\dots,b_n \in \mathbb{R}$,

$$u_0(x) = \sum_{j=1}^{n} b_j \sin(jx). \tag{6.2.6}$$

For this u_0, it can be verified directly that the solution is

$$u(x,t) = \sum_{j=1}^{n} b_j e^{-\nu j^2 t} \sin(jx). \tag{6.2.7}$$

By using the maximum principle for the heat equation (see, e.g. [78] or other textbooks on partial differential equations),

$$\min\{0, \min_{0\le x\le\pi} u_0(x)\} \le u(x,t) \le \max\{0, \max_{0\le x\le\pi} u_0(x)\},$$

we see that

$$\max_{0\le x\le\pi} |u(x,t)| \le \max_{0\le x\le\pi} |u_0(x)| \quad \forall\, t\in[0,T].$$

Thus the solution operator $S(t) : V_0 \subset V \to V$ is bounded.

Then for a general $u_0 \in V$, the problem (6.2.4) has a unique solution. If $u_0 \in V$ has a piecewise continuous derivative in $[0, \pi]$, then from the theory of Fourier series,

$$u_0(x) = \sum_{j=1}^{\infty} b_j \sin(jx)$$

and the solution $u(t)$ can be expressed as

$$u(x,t) = S(t)u_0(x) = \sum_{j=1}^{\infty} b_j e^{-\nu j^2 t} \sin(jx). \qquad \square$$

Return to the abstract problem (6.2.1). We present two results, the first one is on the time continuity of the generalized solution and the second one shows the solution operator $S(t)$ forms a semigroup.

Proposition 6.2.5 *For any $u_0 \in V$, the generalized solution of the initial value problem* (6.2.1) *is continuous in t.*

Proof. Choose a sequence $\{u_{0,n}\} \subset V_0$ that converges to u_0 in V:

$$\|u_{0,n} - u_0\|_V \to 0 \quad \text{as } n \to \infty.$$

Let $t_0 \in [0,T]$ be fixed, and $t \in [0,T]$. We write

$$\begin{aligned} u(t) - u(t_0) &= S(t)u_0 - S(t_0)u_0 \\ &= S(t)(u_0 - u_{0,n}) + [S(t) - S(t_0)]\, u_{0,n} - S(t_0)(u_0 - u_{0,n}). \end{aligned}$$

Then

$$\|u(t) - u(t_0)\|_V \le 2\, c_0 \|u_{0,n} - u_0\|_V + \|\, [S(t) - S(t_0)]\, u_{0,n}\|_V.$$

Given any $\varepsilon > 0$, we choose n sufficiently large such that

$$2\, c_0 \|u_{0,n} - u_0\|_V < \frac{\varepsilon}{2}.$$

For this n, using (6.2.2) of the definition of the solution, we have a $\delta > 0$ such that

$$\|\, [S(t) - S(t_0)]\, u_{0,n}\|_V < \frac{\varepsilon}{2} \quad \text{for } |t - t_0| < \delta.$$

Then for $t \in [0,T]$ with $|t - t_0| < \delta$, we have $\|u(t) - u(t_0)\|_V \le \varepsilon$. $\square$

Proposition 6.2.6 *Assume the problem* (6.2.1) *is well-posed. Then for all $t_1, t_0 \in [0,T]$ such that $t_1 + t_0 \le T$, we have $S(t_1 + t_0) = S(t_1)\, S(t_0)$.*

Proof. The solution of the problem (6.2.1) is $u(t) = S(t)u_0$. We have $u(t_0) = S(t_0)u_0$ and $S(t)u(t_0)$ is the solution of the differential equation on $[t_0, T]$ with the initial condition $u(t_0)$ at t_0. By the uniqueness of the solution,

$$S(t)u(t_0) = u(t + t_0),$$

i.e.,

$$S(t_1)S(t_0)u_0 = S(t_1 + t_0)u_0.$$

Since $u_0 \in V$ is arbitrary, $S(t_1 + t_0) = S(t_1)\, S(t_0)$. □

Now we introduce a finite difference method defined by a one-parameter family of uniformly bounded linear operators

$$C(\Delta t) : V \to V, \quad 0 < \Delta t \leq \Delta_0.$$

Here $\Delta_0 > 0$ is a fixed number. The family $\{C(\Delta t)\}_{0<\Delta t\leq\Delta_0}$ is said to be uniformly bounded if there is a constant c such that

$$\|C(\Delta t)\| \leq c \quad \forall\, \Delta t \in (0, \Delta_0].$$

The approximate solution is then defined by

$$u_{\Delta t}(m\,\Delta t) = C(\Delta t)^m u_0, \quad m = 1, 2, \ldots.$$

Definition 6.2.7 (Consistency) *The difference method is* consistent *if there exists a dense subspace V_c of V such that for all $u_0 \in V_c$, for the corresponding solution u of the initial value problem* (6.2.1)*, we have*

$$\lim_{\Delta t\to 0} \left\| \frac{1}{\Delta t} \left[C(\Delta t)u(t) - u(t + \Delta t)\right] \right\| = 0 \quad \text{uniformly in } [0, T].$$

Assume $V_c \cap V_0 \neq \emptyset$. For $u_0 \in V_c \cap V_0$, we write

$$\begin{aligned} &\frac{1}{\Delta t} \left[C(\Delta t)u(t) - u(t + \Delta t)\right] \\ &\qquad = \left[\frac{C(\Delta t) - I}{\Delta t} - L\right] u(t) - \left[\frac{u(t + \Delta t) - u(t)}{\Delta t} - Lu(t)\right]. \end{aligned}$$

Since

$$\frac{u(t + \Delta t) - u(t)}{\Delta t} - Lu(t) \to 0 \quad \text{as } \Delta t \to 0$$

by the definition of the solution, we see that for a consistent method,

$$\left[\frac{C(\Delta t) - I}{\Delta t} - L\right] u(t) \to 0 \quad \text{as } \Delta t \to 0;$$

so $[C(\Delta t) - I]/\Delta t$ is a convergent approximation of the operator L.

Example 6.2.8 (**continuation of Example 6.2.4**) Let us now consider the forward method and the backward method from Example 6.1.1 for the sample problem (6.2.4). For the forward method, we define the operator $C(\Delta t)$ by the formula

$$C(\Delta t)v(x) = (1-2r)\,v(x) + r\,[v(x+\Delta x) + v(x-\Delta x)]\,,$$

where $\Delta x = \sqrt{\nu \Delta t/r}$ and if $x \pm \Delta x \notin [0,\pi]$, then the function v is extended by oddness with period 2π. We identify Δt with h_t and Δx with h_x. Then $C(\Delta t) : V \to V$ is a linear operator and it can be shown that

$$\|C(\Delta t)v\|_V \le (|1-2r| + 2r)\,\|v\|_V \quad \forall\, v \in V.$$

So

$$\|C(\Delta t)\| \le |1-2r| + 2r, \tag{6.2.8}$$

and the family $\{C(\Delta t)\}$ is uniformly bounded. The difference method is

$$u_{\Delta t}(t_m) = C(\Delta t)u_{\Delta t}(t_{m-1}) = C(\Delta t)^m u_0$$

or

$$u_{\Delta t}(\cdot, t_m) = C(\Delta t)^m u_0(\cdot).$$

Notice that in this form, the difference method generates an approximate solution $u_{\Delta t}(x,t)$ that is defined for $x \in [0,\pi]$ and $t = t_m$, $m = 0, 1, \ldots, N_t$. Since

$$\begin{aligned} u_{\Delta t}(x_j, t_{m+1}) &= (1-2r)\,u_{\Delta t}(x_j, t_m) \\ &\quad + r\,[u_{\Delta t}(x_{j-1}, t_m) + u_{\Delta t}(x_{j+1}, t_m)]\,, \\ &\quad 1 \le j \le N_x - 1,\ 0 \le m \le N_t - 1, \\ u_{\Delta t}(0, t_m) &= u_{\Delta t}(N_x, t_m) = 0, \quad 0 \le m \le N_t, \\ u_{\Delta t}(x_j, 0) &= u_0(x_j), \quad 0 \le j \le N_x, \end{aligned}$$

we see that the relation between the approximate solution $u_{\Delta t}$ and the solution v defined by the ordinary difference scheme (6.1.8)–(6.1.10) (with $f_j^m = 0$) is

$$u_{\Delta t}(x_j, t_m) = v_j^m. \tag{6.2.9}$$

As for the consistency, we take $V_c = V_0$. For the initial value function (6.2.6), we have the formula (6.2.7) for the solution which is obviously

infinitely smooth. Now using Taylor expansions at (x,t), we have

$$\begin{aligned}
&C(\Delta t)u(x,t) - u(x,t+\Delta t)\\
&\quad = (1-2r)u(x,t) + r\,[u(x+\Delta x,t) + u(x-\Delta x,t)] - u(x,t+\Delta t)\\
&\quad = (1-2r)u(x,t) + r\left[2u(x,t) + u_{xx}(x,t)\,(\Delta x)^2\right]\\
&\qquad + \frac{r}{4!}\,[u_{xxxx}(x+\theta_1\Delta x,t) + u_{xxxx}(x-\theta_2\Delta x,t)]\,(\Delta x)^4\\
&\qquad - u(x,t) - u_t(x,t)\,\Delta t - \frac{1}{2}\,u_{tt}(x,t+\theta_3\Delta t)\,(\Delta t)^2\\
&\quad = -\frac{1}{2}\,u_{tt}(x,t+\theta_3\Delta t)\,(\Delta t)^2\\
&\qquad - \frac{\nu^2}{24\,r}\,[u_{xxxx}(x+\theta_1\Delta x,t) + u_{xxxx}(x-\theta_2\Delta x,t)]\,(\Delta t)^2,
\end{aligned}$$

where $\theta_1,\theta_2,\theta_3 \in (0,1)$. Thus,

$$\left\| \frac{1}{\Delta t}\,[C(\Delta t)u(t) - u(t+\Delta t)] \right\| \le c\,\Delta t$$

and we have the consistency of the scheme.

For the backward method, $u_{\Delta t}(\cdot,t+\Delta t) = C(\Delta t)u_{\Delta t}(\cdot,t)$ is defined by

$$\begin{aligned}
(1+2r)\,u_{\Delta t}(x,t+\Delta t) - r\,\big[&u_{\Delta t}(x-\Delta x,t+\Delta t)\\
&+ u_{\Delta t}(x+\Delta x,t+\Delta t)\big] = u_{\Delta t}(x,t)
\end{aligned}$$

with $\Delta x = \sqrt{\nu\Delta t/r}$. Again, for $x \pm \Delta x \notin [0,\pi]$, the function u is extended by oddness with period 2π. Rewrite the relation in the form

$$\begin{aligned}
&u_{\Delta t}(x,t+\Delta t)\\
&\quad = \frac{r}{1+2r}\,[u_{\Delta t}(x-\Delta x,t+\Delta t) + u_{\Delta t}(x+\Delta x,t+\Delta t)] + \frac{u_{\Delta t}(x,t)}{1+2r}.
\end{aligned}$$

Let $\|u_{\Delta t}(\cdot,t+\Delta t)\|_V = |u_{\Delta t}(x_0,t+\Delta t)|$ for some $x_0 \in [0,\pi]$. Then

$$\begin{aligned}
&\|u_{\Delta t}(\cdot,t+\Delta t)\|_V\\
&\quad \le \frac{r}{1+2r}\,[|u_{\Delta t}(x_0-\Delta x,t+\Delta t)| + |u_{\Delta t}(x_0+\Delta x,t+\Delta t)|] + \frac{|u_{\Delta t}(x_0,t)|}{1+2r},
\end{aligned}$$

i.e.,

$$\|u_{\Delta t}(\cdot,t+\Delta t)\|_V \le \frac{2r}{1+2r}\|u_{\Delta t}(\cdot,t+\Delta t)\|_V + \frac{\|u_{\Delta t}(\cdot,t)\|_V}{1+2r}.$$

So

$$\|u_{\Delta t}(\cdot,t+\Delta t)\|_V \le \|u_{\Delta t}(\cdot,t)\|_V$$

and the family $\{C(\Delta t)\}_{0<\Delta t\le\Delta_0}$ is uniformly bounded for any fixed small number $\Delta_0 > 0$.

Showing consistency of the backward scheme is more involved, and the argument is similar to that in Example 6.3.4 where the definition of the consistency is slightly different but is essentially the same. □

Let us return to the general case.

Definition 6.2.9 (CONVERGENCE) *The difference method is* convergent *if for any fixed $t \in [0,T]$, any $u_0 \in V$, we have*

$$\lim_{\Delta t_i \to 0} \| [C(\Delta t_i)^{m_i} - S(t)]\, u_0 \| = 0$$

where $\{m_i\}$ is a sequence of integers and $\{\Delta t_i\}$ a sequence of step sizes such that $\lim_{i\to\infty} m_i \Delta t_i = t$.

Definition 6.2.10 (STABILITY) *The difference method is* stable *if the operators*

$$\{C(\Delta t)^m \mid 0 < \Delta t \le \Delta_0,\ m\Delta t \le T\}$$

are uniformly bounded, i.e., there exists a constant $M_0 > 0$ such that

$$\|C(\Delta t)^m\|_{V\to V} \le M_0 \quad \forall\, m : m\Delta t \le T,\ \forall\, \Delta t \le \Delta_0.$$

We now come to the central result of the section.

Theorem 6.2.11 (LAX EQUIVALENCE THEOREM) *Assume the initial value problem* (6.2.1) *is well-posed. Then, for a consistent difference method, stability is equivalent to convergence.*

Proof. ($\Longrightarrow$) Consider the error

$$\begin{aligned} &C(\Delta t)^m u_0 - u(t) \\ &\quad = \sum_{j=1}^{m-1} C(\Delta t)^j [C(\Delta t) u((m-1-j)\Delta t) - u((m-j)\Delta t)] \\ &\qquad + u(m\Delta t) - u(t). \end{aligned}$$

First assume $u_0 \in V_c$. Then since the method is stable,

$$\begin{aligned} \|C(\Delta t)^m u_0 - u(t)\| \le M_0\, m\Delta t \sup_t \left\| \frac{C(\Delta t)u(t) - u(t+\Delta t)}{\Delta t} \right\| \\ + \|u(m\Delta t) - u(t)\|. \end{aligned} \tag{6.2.10}$$

By continuity, $\|u(m\Delta t) - u(t)\| \to 0$, and by the consistency,

$$\sup_t \left\| \frac{C(\Delta t)u(t) - u(t+\Delta t)}{\Delta t} \right\| \to 0.$$

So we have the convergence by (6.2.10).

Next consider the convergence for the general case where $u_0 \in V$. We have a sequence $\{u_{0,n}\} \subset V_0$ such that $u_{0,n} \to u_0$ in V. Writing

$$\begin{aligned} C(\Delta t)^m u_0 - u(t) = C(\Delta t)^m (u_0 - u_{0,n}) \\ + [C(\Delta t)^m - S(t)]\, u_{0,n} - S(t)\,(u_0 - u_{0,n}), \end{aligned}$$

we obtain

$$\|C(\Delta t)^m u_0 - u(t)\| \le \|C(\Delta t)^m (u_0 - u_{0,n})\| + \|[C(\Delta t)^m - S(t)]u_{0,n}\| + \|S(t)\,(u_0 - u_{0,n})\|.$$

Since the initial value problem (6.2.1) is well-posed and the method is stable,

$$\|C(\Delta t)^m u_0 - u(t)\| \le c\,\|u_0 - u_{0,n}\| + \|[C(\Delta t)^m - S(t)]u_{0,n}\|.$$

Given any $\varepsilon > 0$, there is an n sufficiently large such that

$$c\,\|u_0 - u_{0,n}\| < \frac{\varepsilon}{2}.$$

For this n, let Δt be sufficiently small,

$$\|[C(\Delta t)^m - S(t)]u_{0,n}\| < \frac{\varepsilon}{2} \quad \forall\, \Delta t \text{ small},\ |m\Delta t - t| < \Delta t.$$

Then we obtain the convergence.

($\Longleftarrow$) Suppose the method is not stable. Then there are sequences $\{\Delta t_k\}$ and $\{m_k\}$ such that $m_k \Delta t_k \le T$ and

$$\lim_{k\to\infty} \|C(\Delta t_k)^{m_k}\| = \infty.$$

Since $\Delta t_k \le \Delta_0$, we may assume the sequence $\{\Delta t_k\}$ is convergent. If the sequence $\{m_k\}$ is bounded, then

$$\sup_k \|C(\Delta t_k)^{m_k}\| \le \sup_k \|C(\Delta t_k)\|^{m_k} < \infty.$$

This is a contradiction. Thus $m_k \to \infty$ and $\Delta t_k \to 0$ as $k \to \infty$.

By the convergence of the method,

$$\sup_k \|C(\Delta t_k)^{m_k} u_0\| < \infty \quad \forall\, u_0 \in V.$$

Applying Theorem 2.4.4, we have

$$\lim_{k\to\infty} \|C(\Delta t_k)^{m_k}\| < \infty,$$

contradicting the assumption that the method is not stable. □

Corollary 6.2.12 (Convergence order) *Under the assumptions of Theorem 6.2.11, if u is a solution with initial value $u_0 \in V_c$ satisfying*

$$\sup_{0\le t\le T} \left\| \frac{C(\Delta t)u(t) - u(t+\Delta t)}{\Delta t} \right\| \le c\,(\Delta t)^k \quad \forall\, \Delta t \in (0, \Delta_0],$$

then we have the error estimate

$$\|C(\Delta t)^m u_0 - u(t)\| \le c\,(\Delta t)^k,$$

where m is a positive integer with $m\Delta t = t$.

Proof. The error estimate follows immediately from (6.2.10). □

Example 6.2.13 (**continuation of Example 6.2.8**) Let us apply the Lax equivalence theorem to the forward and backward schemes for the sample problem 6.2.4. For the forward scheme, we assume $r \leq 1/2$. Then according to (6.2.8), $\|C(\Delta t)\| \leq 1$ and so

$$\|C(\Delta t)^m\| \leq 1, \quad m = 1, 2, \ldots.$$

Thus under the condition $r \leq 1/2$, the forward scheme is stable. Since the scheme is consistent, we have the convergence

$$\lim_{\Delta t_i \to 0} \|u_{\Delta t}(\cdot, m_i \Delta t_i) - u(\cdot, t)\|_V = 0, \tag{6.2.11}$$

where $\lim_{\Delta t_i \to 0} m_i \Delta t_i = t$.

Actually, it can be shown that

$$\|C(\Delta t)\| = |1 - 2r| + 2r$$

and $r \leq 1/2$ is a necessary and sufficient condition for stability and then for convergence as well (Exercise 6.2.3).

By the relation (6.2.9), for the finite difference solution $\{v_j^m\}$ defined in (6.1.8)–(6.1.10) with $f_j^m = 0$, we have the convergence

$$\lim_{h_t \to 0} \max_{0 \leq j \leq N_x} |v_j^m - u(x_j, t)| = 0,$$

where m depends on h_t and $\lim_{h_t \to 0} m h_t = t$.

Since we need a condition ($r \leq 1/2$ in this case) for convergence, the forward scheme is said to be *conditionally stable* and *conditionally convergent*.

For the backward scheme, for any r, $\|C(\Delta t)\| \leq 1$. Then

$$\|C(\Delta t)^m\| \leq 1 \quad \forall\, m.$$

So the backward scheme is *unconditionally stable*, which leads to *unconditional convergence* of the backward scheme. We skip the detailed presentation of the arguments for the above statement since the arguments are similar to those for the forward scheme.

We can also apply Corollary 6.2.12 to claim convergence order for the forward and backward schemes, see Examples 6.3.3 and 6.3.4 in the next section for some similar arguments. □

Exercise 6.2.1 Show that the subspace V_0 defined in (6.2.5) is dense in V.

Exercise 6.2.2 Analyze the Crank-Nicolson scheme for the problem (6.2.4).

Exercise 6.2.3 Consider the forward scheme for solving the sample problem 6.2.4. Show that

$$\|C(\Delta t)\| = |1 - 2r| + 2r,$$

and $r \leq 1/2$ is a necessary and sufficient condition for both stability and convergence.

6.3 More on convergence

In the literature, one can find various slightly different variants of the Lax equivalence theorem presented in the preceding section. Here we consider one such variant which is usually more convenient to apply in analyzing convergence of difference schemes for solving initial-boundary value problems.

Consider an initial-boundary value problem of the form

$$Lu = f \quad \text{in } (0,a) \times (0,T), \tag{6.3.1}$$

$$u(0,t) = u(a,t) = 0, \quad t \in [0,T], \tag{6.3.2}$$

$$u(x,0) = u_0(x), \quad x \in [0,a]. \tag{6.3.3}$$

Here f and u_0 are given continuous functions, and L is a linear partial differential operator of first order with respect to the time variable. For the problem (6.1.5)–(6.1.7), $L = \partial_t - \nu\,\partial_x^2$. We assume for the given data f and u_0, the problem (6.3.1)–(6.3.3) has a unique solution u with certain smoothness that makes the following calculations meaningful (e.g., derivatives of u up to certain order are continuous).

Again denote N_x and N_t positive integers, $h_x = a/N_x$, $h_t = T/N_t$, and we use the other notations introduced in Example 6.1.1. Corresponding to the time level $t = t_m$, we introduce the solution vector

$$\boldsymbol{v}^m = (v_1^m, \ldots, v_{N_x-1}^m)^T \in \mathbb{R}^{N_x-1},$$

where the norm in the space $\mathbb{R}^{N_x-1}$ is denoted by $\|\cdot\|$; this norm depends on the dimension $N_x - 1$, but we do not indicate this dependence explicitly for notational simplicity. We will be specific about the norm when we consider concrete examples.

Consider a general two-level scheme

$$\boldsymbol{v}^{m+1} = Q\boldsymbol{v}^m + h_t\boldsymbol{g}^m, \quad 0 \leq m \leq N_t - 1, \tag{6.3.4}$$

$$\boldsymbol{v}^0 = \boldsymbol{u}^0. \tag{6.3.5}$$

Here the matrix $Q \in \mathbb{R}^{(N_x-1)\times(N_x-1)}$ may depend on h_t and h_x. We use $\|Q\|$ to denote the operator matrix norm induced by the vector norm on $\mathbb{R}^{N_x-1}$. The vector $\boldsymbol{g}^m$ is usually constructed from values of f at $t = t_m$,

$$\boldsymbol{u}^0 = (u_0(x_1), \ldots, u_0(x_{N_x-1}))^T,$$

and in general

$$\boldsymbol{u}^m = (u_1^m, \ldots, u_{N_x-1}^m)^T, \quad 1 \le m \le N_t,$$

with u the solution of (6.3.1)–(6.3.3).

We now introduce definitions of consistency, stability and convergence for the scheme (6.3.4)–(6.3.5). For this purpose, we need to define a quantity $\boldsymbol{\tau}^m$ through the relation

$$\boldsymbol{u}^{m+1} = Q\boldsymbol{u}^m + h_t\boldsymbol{g}^m + h_t\boldsymbol{\tau}^m. \tag{6.3.6}$$

This quantity $\boldsymbol{\tau}^m$ can be called the *local truncation error* of the scheme. As we will see from examples below, for an explicit method, $\boldsymbol{\tau}^m$ defined in (6.3.6) is indeed the local truncation error used in many references. In the case of an implicit method, $\boldsymbol{\tau}^m$ defined here is related to the usual local truncation error by a linear transformation.

Definition 6.3.1 *We say the difference method* (6.3.4)–(6.3.5) *is* consistent *if*

$$\sup_{m:mh_t\le T} \|\boldsymbol{\tau}^m\| \to 0 \quad \text{as } h_t, h_x \to 0.$$

The method is of order (p_1, p_2) *if, when the solution* u *is sufficiently smooth, there is a constant* c *such that*

$$\sup_{m:mh_t\le T} \|\boldsymbol{\tau}^m\| \le c\,(h_x^{p_1} + h_t^{p_2}). \tag{6.3.7}$$

The method is said to be stable *if for some constant* $M_0 < \infty$*, which may depend on* T*, we have*

$$\sup_{m:mh_t\le T} \|Q^m\| \le M_0.$$

The method is convergent *if*

$$\sup_{m:mh_t\le T} \|\boldsymbol{u}^m - \boldsymbol{v}^m\| \to 0 \quad \text{as } h_t, h_x \to 0.$$

We have the following theorem concerning convergence and convergence order of the difference method.

Theorem 6.3.2 *Assume the scheme* (6.3.4)–(6.3.5) *is consistent and stable. Then the method is convergent. Moreover, if the solution* u *is sufficiently smooth so that* (6.3.7) *holds, then we have the error estimate*

$$\sup_{m:mh_t\le T} \|\boldsymbol{u}^m - \boldsymbol{v}^m\| \le c\,(h_x^{p_1} + h_t^{p_2}).$$

Proof. Introduce the error vectors: $\boldsymbol{e}^m = \boldsymbol{u}^m - \boldsymbol{v}^m$ for $m = 0, 1, \ldots, N_t$. Then $\boldsymbol{e}^0 = \boldsymbol{0}$ by the definition of the initial value for the scheme. We have the error recursion relation, derived from (6.3.4) and (6.3.6):

$$\boldsymbol{e}^{m+1} = Q\,\boldsymbol{e}^m + h_t\boldsymbol{\tau}^m.$$

Using this relation repeatedly and remembering $\boldsymbol{e}^0 = \boldsymbol{0}$, we find

$$\boldsymbol{e}^{m+1} = h_t \sum_{l=0}^{m} Q^l \boldsymbol{\tau}^{m-l}.$$

Thus

$$\|\boldsymbol{e}^{m+1}\| \le h_t \sum_{l=0}^{m} \|Q^l\|\,\|\boldsymbol{\tau}^{m-l}\|.$$

Apply the stability condition,

$$\|\boldsymbol{e}^{m+1}\| \le M_0(m+1)\,h_t \sup_{0\le l\le m} \|\boldsymbol{\tau}^{m-l}\|.$$

Then we have the inequality

$$\sup_{m:mh_t\le T} \|\boldsymbol{u}^m - \boldsymbol{v}^m\| \le M_0 T \sup_{m:mh_t\le T} \|\boldsymbol{\tau}^m\|,$$

and the claims of the theorem follow. □

Example 6.3.3 We give a convergence analysis of the scheme (6.1.8)–(6.1.10). Assume $u_{tt}, u_{xxxx} \in C([0,\pi]\times[0,T])$. Then (cf. (6.1.11))

$$u_j^{m+1} = (1-2r)\,u_j^m + r\,(u_{j+1}^m + u_{j-1}^m) + h_t f_j^m + h_t\tau_j^m,$$

where, following easily from Taylor expansions,

$$|\tau_j^m| \le c\,(h_x^2 + h_t)$$

with the constant c depending on u_{tt} and u_{xxxx}.

The scheme (6.1.8)–(6.1.10) can be written in the form (6.3.4)–(6.3.5) with

$$Q = \begin{pmatrix} 1-2r & r & & & \\ r & 1-2r & r & & \\ & \ddots & \ddots & \ddots & \\ & & r & 1-2r & r \\ & & & r & 1-2r \end{pmatrix}_{(N_x-1)\times(N_x-1)}$$

and

$$\boldsymbol{g}^m = \boldsymbol{f}^m \equiv (f_1^m, \ldots, f_{N_x-1}^m)^T.$$

Let us assume the condition $r \le 1/2$. Then if we choose to use the maximum-norm, we have

$$\|Q\|_\infty = 1, \quad \|\boldsymbol{\tau}^m\|_\infty \le c\,(h_x^2 + h_t).$$

Thus the method is stable, and we can apply Theorem 6.3.2 to conclude that under the conditions $u_{tt}, u_{xxxx} \in C([0,\pi]\times[0,T])$ and $r \le 1/2$,

$$\max_{0\le m\le N_x} \|\boldsymbol{u}^m - \boldsymbol{v}^m\|_\infty \le c\,(h_x^2 + h_t).$$

Now suppose we use the discrete weighted 2-norm:

$$\|\boldsymbol{v}\|_{2,h_x} = \sqrt{h_x}\|\boldsymbol{v}\|_2.$$

It is easy to see that the induced matrix norm is the usual spectral norm $\|Q\|_2$. Since Q is symmetric and its eigenvalues are (see Exercise 6.3.1)

$$\lambda_j(Q) = 1 - 4r\,\sin^2\Big(\frac{j\pi}{2N_x}\Big), \quad 1 \le j \le N_x - 1,$$

we see that under the condition $r \le 1/2$,

$$\|Q\|_2 = \max_j |\lambda_j(Q)| < 1,$$

i.e., the method is stable. It is easy to verify that

$$\|\boldsymbol{\tau}^m\|_{2,h_x} \le c\,(h_x^2 + h_t).$$

Thus by Theorem 6.3.2, we conclude that under the conditions $u_{tt}, u_{xxxx} \in C([0,\pi]\times[0,T])$ and $r \le 1/2$,

$$\max_{0\le m\le N_x} \|\boldsymbol{u}^m - \boldsymbol{v}^m\|_{2,h_x} \le c\,(h_x^2 + h_t).$$

So the convergence order is 2 in h_x and 1 in h_t. □

Example 6.3.4 Now consider the backward scheme (6.1.12)–(6.1.14). Assume $u_{tt}, u_{xxxx} \in C([0,\pi]\times[0,T])$. Then (cf. (6.1.15))

$$(1+2r)\,u_j^m - r\,(u_{j+1}^m + u_{j-1}^m) = u_j^{m-1} + h_t f_j^m + h_t\overline{\tau}_j^m,$$

where

$$|\overline{\tau}_j^m| \le c\,(h_x^2 + h_t)$$

with the constant c depending on u_{tt} and u_{xxxx}. Define the matrix

$$Q = Q_1^{-1}$$

where

$$Q_1 = \begin{pmatrix} 1+2r & -r & & & \\ -r & 1+2r & -r & & \\ & \ddots & \ddots & \ddots & \\ & & -r & 1+2r & -r \\ & & & -r & 1+2r \end{pmatrix}_{(N_x-1)\times(N_x-1)}.$$

Let $\boldsymbol{g}^m = Q\boldsymbol{f}^m$ and $\boldsymbol{\tau}^m = Q\overline{\boldsymbol{\tau}}^m$. Then the scheme (6.1.12)–(6.1.14) can be written in the form (6.3.4)–(6.3.5).

First we consider the convergence in $\|\cdot\|_\infty$. Let us estimate $\|Q\|_\infty$. From the definition of Q,

$$\boldsymbol{y} = Q\boldsymbol{x} \Longleftrightarrow \boldsymbol{x} = Q_1\boldsymbol{y} \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{N_x-1}.$$

Thus

$$y_i = \frac{r}{1+2r}(y_{i-1} + y_{i+1}) + \frac{x_i}{1+2r}, \quad 1 \le i \le N_x - 1.$$

Suppose $\|\boldsymbol{y}\|_\infty = |y_i|$. Then

$$\|\boldsymbol{y}\|_\infty = |y_i| \le \frac{r}{1+2r}\, 2\, \|\boldsymbol{y}\|_\infty + \frac{\|\boldsymbol{x}\|_\infty}{1+2r}.$$

So

$$\|Q\boldsymbol{x}\|_\infty = \|\boldsymbol{y}\|_\infty \le \|\boldsymbol{x}\|_\infty \quad \forall\, \boldsymbol{x} \in \mathbb{R}^{N_x-1}.$$

Hence

$$\|Q\|_\infty \le 1,$$

the backward scheme is unconditionally stable and it is easy to see

$$\|\boldsymbol{\tau}^m\|_\infty \le \|\overline{\boldsymbol{\tau}}^m\|_\infty.$$

Applying Theorem 6.3.2, for the backward scheme (6.1.12)–(6.1.14), we conclude that under the conditions $u_{tt}, u_{xxxx} \in C([0,\pi]\times[0,T])$,

$$\max_{0\le m\le N_x} \|\boldsymbol{u}^m - \boldsymbol{v}^m\|_\infty \le c\,(h_x^2 + h_t).$$

Now we consider the convergence in $\|\cdot\|_{2,h_x}$. By Exercise 6.3.1, the eigenvalues of Q_1 are

$$\lambda_j(Q_1) = 1 + 4r\,\cos^2\frac{j\pi}{2N_x}, \quad 1 \le j \le N_x - 1.$$

Since $Q = Q_1^{-1}$, the eigenvalues of Q are

$$\lambda_j(Q) = \lambda_j(Q_1)^{-1} \in (0,1), \quad 1 \le j \le N_x - 1.$$

Now that Q is symmetric because Q_1 is,

$$\|Q\|_2 = \max_{1\le j\le N_x-1} |\lambda_j(Q)| < 1.$$

So the backward scheme is unconditionally stable measured in $\|\cdot\|_{2,h_x}$, and it is also easy to deduce

$$\|\boldsymbol{\tau}^m\|_{2,h_x} \le \|\overline{\boldsymbol{\tau}}^m\|_{2,h_x}.$$

So for the backward scheme (6.1.12)–(6.1.14), we apply Theorem 6.3.2 to conclude that under the conditions $u_{tt}, u_{xxxx} \in C([0,\pi]\times[0,T])$,

$$\max_{0\le m\le N_x} \|\boldsymbol{u}^m - \boldsymbol{v}^m\|_{2,h_x} \le c\,(h_x^2 + h_t).$$

Again, the convergence order is 2 in h_x and 1 in h_t. □

Exercise 6.3.1 Let $a, b, c \in \mathbb{R}$ with $bc \ge 0$. Show that the eigenvalues of the matrix

$$Q = \begin{pmatrix} a & c & & & \\ b & a & c & & \\ & \ddots & \ddots & \ddots & \\ & & b & a & c \\ & & & b & a \end{pmatrix}_{N\times N}$$

are

$$\lambda_j = a + 2\sqrt{bc}\,\cos\Big(\frac{j\pi}{N+1}\Big), \quad 1 \le j \le N.$$

Hint: For the nontrivial case $bc \ne 0$, write

$$Q = a\,I + \sqrt{bc}\,D^{-1}\Lambda D$$

with D is a diagonal matrix with the diagonal elements $\sqrt{c/b}$, $(\sqrt{c/b})^2$, …, $(\sqrt{c/b})^N$, and Λ is a tridiagonal matrix

$$\Lambda = \begin{pmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & & 1 & 0 \end{pmatrix}.$$

Then find the eigenvalues of Λ by following the definition of the eigenvalue problem and solving a difference system for components of associated eigenvectors. An alternative approach is to relate the characteristic equation of Λ through its recursion formula to Chebyshev polynomials of the second kind ([15, p. 497]).

Exercise 6.3.2 Provide a convergence analysis for the Crank-Nicolson scheme (6.1.16)–(6.1.18) in $\|\cdot\|_{2,h_x}$ norm by applying Theorem 6.3.2, as is done for the forward and backward schemes in examples.

Exercise 6.3.3 The forward, backward and Crank-Nicolson schemes are all particular members in a family of difference schemes called *generalized mid-point methods*. Let $\theta \in [0,1]$ be a parameter. Then a generalized mid-point scheme for the initial-boundary value problem (6.1.5)–(6.1.7) is

$$\frac{v_j^m - v_j^{m-1}}{h_t} = \nu\,\theta\,\frac{v_{j+1}^m - 2v_j^m + v_{j-1}^m}{h_x^2} + \nu\,(1-\theta)\,\frac{v_{j+1}^{m-1} - 2v_j^{m-1} + v_{j-1}^{m-1}}{h_x^2} + \theta\,f_j^m + (1-\theta)\,f_j^{m-1}, \quad 1 \le j \le N_x - 1,\ 1 \le m \le N_t,$$

supplemented by the boundary condition (6.1.13) and the initial condition (6.1.14). Show that for $\theta \in [1/2, 1]$, the scheme is unconditionally stable in both $\|\cdot\|_{2,h_x}$ and $\|\cdot\|_\infty$ norms; for $\theta \in [0, 1/2)$, the scheme is stable in $\|\cdot\|_{2,h_x}$ norm if $2\,(1-2\,\theta)\,r \le 1$, and it is stable in $\|\cdot\|_\infty$ norm if $2\,(1-\theta)\,r \le 1$. Determine the convergence orders of the schemes.

Suggestion for Further Reading.

More details on theoretical analysis of the finite difference method, e.g., treatment of other kind of boundary conditions, general spatial domains for higher spatial dimension problems, approximation of hyperbolic or elliptic problems, can be found in several books on the topic, e.g., [218]. For the finite difference method for parabolic problems, [221] is an in-depth survey. Another popular approach to developing finite difference methods for parabolic problems is the *method of lines*; see [15, p. 414] for an introduction which discusses some of the finite difference methods of this chapter. The book [101] provides a comprehensive coverage of high order finite difference methods for time dependent PDEs.

For initial-boundary value problems of evolution equations in high spatial dimension, stability for an explicit scheme usually requires the time stepsize to be prohibitively small. On the other hand, some implicit schemes are unconditionally stable, and stability requirement does not impose restriction on the time stepsize. The disadvantage of an implicit scheme is that at each time level we may need to solve an algebraic system of very large scale. The idea of *operator splitting technique* is to split the computation for each time step into several substeps such that each substep is implicit only in one spatial variable and at the same time good stability property is maintained. The resulting schemes are called *alternating direction methods* or *fractional step methods*. See [165, 243] for detailed discussions.

Many physical phenomena are described by conservation laws (conservation of mass, momentum, and energy). Finite difference methods for conservation laws constitute a large research area. The interested reader can consult [155] and [94].

Extrapolation methods are efficient means to accelerate the convergence of numerical solutions. For extrapolation methods in the context of the finite difference method, see [166].

7
Sobolev Spaces

In this chapter, we review definitions and properties of Sobolev spaces, which are indispensable for a theoretical analysis of partial differential equations and boundary integral equations, as well as being necessary for the analysis of some numerical methods for solving such equations. Most results are stated without proof; proofs of the results and detailed discussions of Sobolev spaces can be found in numerous monographs and textbooks, e.g. [1, 78, 89, 245].

7.1 Weak derivatives

We need the multi-index notation for partial derivatives introduced in Section 1.4.

Our purpose in this section is to extend the definition of derivatives. To do this, we start with the classical "integration by parts" formula

$$\int_\Omega v(\boldsymbol{x})\,\partial^\alpha \phi(\boldsymbol{x})\,dx = (-1)^{|\alpha|}\int_\Omega \partial^\alpha v(\boldsymbol{x})\,\phi(\boldsymbol{x})\,dx \tag{7.1.1}$$

which holds for $v \in C^m(\Omega)$, $\phi \in C_0^\infty(\Omega)$ and $|\alpha| \le m$. This formula, relating differentiation and integration, is a most important formula in calculus. The *weak derivative* is defined in such a way that, first, if the classical derivative exists then the two derivatives coincide so that the weak derivative is an extension of the classical derivative; second, the integration by parts formula (7.1.1) holds, where ∂^α is now a weak differential operator applied

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_7,
© Springer Science + Business Media, LLC 2009

to less smooth functions. A more general approach for the extension of the classical derivatives is to first introduce the derivatives in the distributional sense. A detailed discussion of distributions and the derivatives in the distributional sense can be found in several monographs, e.g., [210]. Here we choose to introduce the concept of the weak derivatives directly, which is sufficient for this text.

As preparation, we first introduce the notion of locally integrable functions.

Definition 7.1.1 *Let $1 \le p < \infty$. A function $v : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ is said to be* locally p-integrable, *$v \in L^p_{\rm loc}(\Omega)$, if for every $\boldsymbol{x} \in \Omega$, there is an open neighborhood Ω' of $\boldsymbol{x}$ such that $\overline{\Omega'} \subset \Omega$ and $v \in L^p(\Omega')$.*

Notice that a locally integrable function can behave arbitrarily badly near the boundary $\partial\Omega$. One such example is the function $e^{d(\boldsymbol{x})^{-1}} \sin(d(\boldsymbol{x})^{-1})$, where $d(\boldsymbol{x}) \equiv \inf_{\boldsymbol{y}\in\partial\Omega} \|\boldsymbol{x} - \boldsymbol{y}\|$ is the distance from $\boldsymbol{x}$ to $\partial\Omega$.

We have the following useful result which will be used repeatedly ([245, p. 18]).

Lemma 7.1.2 (Generalized Variational Lemma) *Let $v \in L^1_{\rm loc}(\Omega)$ with Ω a nonempty open set in $\mathbb{R}^d$. If*

$$\int_\Omega v(\boldsymbol{x})\,\phi(\boldsymbol{x})\,dx = 0 \quad \forall\, \phi \in C_0^\infty(\Omega),$$

then $v = 0$ a.e. on Ω.

Now we are ready to introduce the concept of a weak derivative.

Definition 7.1.3 *Let Ω be a nonempty open set in $\mathbb{R}^d$, $v, w \in L^1_{\rm loc}(\Omega)$. Then w is called a* weak α^{th} derivative *of v if*

$$\int_\Omega v(\boldsymbol{x})\,\partial^\alpha\phi(\boldsymbol{x})\,dx = (-1)^{|\alpha|}\int_\Omega w(\boldsymbol{x})\,\phi(\boldsymbol{x})\,dx \quad \forall\, \phi \in C_0^\infty(\Omega). \tag{7.1.2}$$

We also say w is a weak derivative of v of order $|\alpha|$.

Lemma 7.1.4 *A weak derivative, if it exists, is uniquely defined up to a set of measure zero.*

Proof. Suppose $v \in L^1_{\rm loc}(\Omega)$ has two weak α^{th} derivatives $w_1, w_2 \in L^1_{\rm loc}(\Omega)$. Then from the definition, we have

$$\int_\Omega [w_1(\boldsymbol{x}) - w_2(\boldsymbol{x})]\,\phi(\boldsymbol{x})\,dx = 0 \quad \forall\, \phi \in C_0^\infty(\Omega).$$

Applying Lemma 7.1.2, we conclude $w_1 = w_2$ a.e. on Ω. □

From the definition of the weak derivative and Lemma 7.1.4, we immediately see the following result holds.

Lemma 7.1.5 *If $v \in C^m(\Omega)$, then for each α with $|\alpha| \le m$, the classical partial derivative $\partial^\alpha v$ is also the weak α^{th} partial derivative of v.*

Because of Lemma 7.1.5, it is natural to use all those notations of the classical derivatives also for the weak derivatives. For example, $\partial_i v \equiv v_{x_i}$ denote the first-order weak derivative of v with respect to x_i.

The weak derivatives defined here coincide with the extension of the classical derivatives discussed in Section 2.4. Let us return to the situation of Example 2.4.2, where the classical differentiation operator $D : C^1[0,1] \to L^2(0,1)$ is extended to the differentiation operator $\hat{D}$ defined over $H^1(0,1)$, the completion of $C^1[0,1]$ under the norm $\|\cdot\|_{1,2}$. For any $v \in H^1(0,1)$, there exists a sequence $\{v_n\} \subset C^1[0,1]$ such that

$$\|v_n - v\|_{1,2} \to 0 \quad \text{as } n \to \infty,$$

which implies, as $n \to \infty$,

$$v_n \to v \quad \text{and} \quad Dv_n \to \hat{D}v \quad \text{in } L^2(0,1).$$

Now by the classical integration by parts formula,

$$\int_0^1 v_n(x)\, D\phi(x)\, dx = -\int_0^1 Dv_n(x)\, \phi(x)\, dx \quad \forall\, \phi \in C_0^\infty(0,1).$$

Taking the limit as $n \to \infty$ in the above relation, we obtain

$$\int_0^1 v(x)\, D\phi(x)\, dx = -\int_0^1 \hat{D}v(x)\, \phi(x)\, dx \quad \forall\, \phi \in C_0^\infty(0,1).$$

Hence, $\hat{D}v$ is also the first-order weak derivative of v.

Now we examine some examples of weakly differentiable functions which are not differentiable in the classical sense, as well as some examples of functions which are not weakly differentiable.

Example 7.1.6 The absolute value function $v(x) = |x|$ is not differentiable at $x = 0$ in the classical sense. Nevertheless, the function is weakly differentiable on any interval containing the point $x = 0$. Indeed, it is easy to verify that

$$w(x) = \begin{cases} 1, & x > 0, \\ -1, & x < 0, \\ c_0, & x = 0, \end{cases}$$

where $c_0 \in \mathbb{R}$ is arbitrary, is a first order weak derivative of the absolute value function. □

Example 7.1.7 Functions with jump discontinuities are not weakly differentiable. For example, define

$$v(x) = \begin{cases} -1, & -1 < x < 0, \\ c_0, & x = 0, \\ 1, & 0 < x < 1, \end{cases}$$

where $c_0 \in \mathbb{R}$. Let us show that the function v does not have a weak derivative. We argue by contradiction. Suppose v is weakly differentiable with the derivative $w \in L^1_{\text{loc}}(-1,1)$. By definition, we have the identity

$$\int_{-1}^{1} v(x)\,\phi'(x)\,dx = -\int_{-1}^{1} w(x)\,\phi(x)\,dx \quad \forall\, \phi \in C_0^\infty(-1,1).$$

The left hand side of the relation can be simplified to $-2\,\phi(0)$. Hence we have the identity

$$\int_{-1}^{1} w(x)\,\phi(x)\,dx = 2\,\phi(0) \quad \forall\, \phi \in C_0^\infty(-1,1).$$

Taking $\phi \in C_0^\infty(0,1)$, viewed as a function in $C_0^\infty(-1,1)$ with vanishing value on $[-1,0]$, we get

$$\int_{0}^{1} w(x)\,\phi(x)\,dx = 0 \quad \forall\, \phi \in C_0^\infty(0,1).$$

By Lemma 7.1.2, we conclude that $w(x) = 0$ a.e. on $(0,1)$. Similarly, $w(x) = 0$ a.e. on $(-1,0)$. So $w(x) = 0$ a.e. on $(-1,1)$, and we arrive at the contradictory relation

$$0 = 2\,\phi(0) \quad \forall\, \phi \in C_0^\infty(-1,1).$$

Thus the function v is not weakly differentiable. □

Example 7.1.8 More generally, assume $v \in C[a,b]$ is piecewisely continuously differentiable (Figure 7.1), i.e., there exist a partition of the interval: $a = x_0 < x_1 < \cdots < x_n = b$, such that $v \in C^1[x_{i-1}, x_i]$, $1 \le i \le n$. Then the first-order weak derivative of v is

$$w(x) = \begin{cases} v'(x), & x \in \cup_{i=1}^{n}(x_{i-1}, x_i), \\ \text{arbitrary}, & x = x_i, \quad 0 \le i \le n. \end{cases}$$

This result can be verified directly by applying the definition of the weak derivative. Suppose moreover that v is piecewise smoother, e.g., $v \in C^2[x_{i-1}, x_i]$, $1 \le i \le n$. Similar to Example 7.1.7, we can show that a second-order weak derivative of v does not exist, unless $v'(x_i-) = v'(x_i+)$ for $i = 1, \ldots, n-1$. □

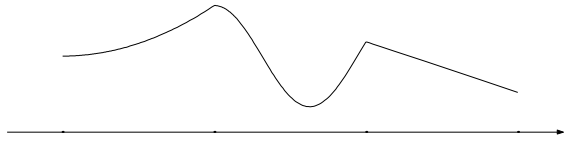

FIGURE 7.1. A continuous function that is piecewisely smooth

Example 7.1.9 In the finite element analysis for solving differential and integral equations, we frequently deal with piecewise polynomials, or piecewise images of polynomials. Suppose $\Omega \subset \mathbb{R}^2$ is a polygonal domain and is partitioned into polygonal sub-domains:

$$\overline{\Omega} = \bigcup_{n=1}^{N} \overline{\Omega_n}.$$

Each sub-domain Ω_n is usually taken to be a triangle or a quadrilateral. Suppose for some non-negative integer k,

$$v \in C^k(\overline{\Omega}), \quad v|_{\Omega_n} \in C^{k+1}(\overline{\Omega_n}),\ 1 \le n \le N.$$

Then the weak partial derivatives of v of order $(k+1)$ exist, and for a multi-index α of length $(k+1)$, the α^{th} weak derivative of v is given by the formula

$$\begin{cases} \partial^\alpha v(\boldsymbol{x}), & \boldsymbol{x} \in \cup_{n=1}^{N} \Omega_n, \\ \text{arbitrary}, & \text{otherwise.} \end{cases}$$

When Ω is a general curved domain, Ω_n may be a curved triangle or quadrilateral. Finite element functions are piecewise polynomials or images of piecewise polynomials. The index k is determined by the order of the PDE problem and the type of the finite elements (conforming or non-conforming). For example, for a second-order elliptic boundary value problem, finite element functions for a conforming method are globally continuous and have first-order weak derivatives. For a non-conforming method, the finite element functions are not globally continuous, and hence do not have first-order weak derivatives. Nevertheless, such functions are smooth in each sub-domain (element). For details, see Chapter 10. □

Most differentiation rules for classical derivatives can be carried over to weak derivatives. Two such examples are the following results.

Proposition 7.1.10 *Let α be a multi-index, $c_1, c_2 \in \mathbb{R}$. If $\partial^\alpha u$ and $\partial^\alpha v$ exist, then so does $\partial^\alpha(c_1 u + c_2 v)$ and*

$$\partial^\alpha(c_1 u + c_2 v) = c_1 \partial^\alpha u + c_2 \partial^\alpha v.$$

Proposition 7.1.11 *Let $p, q \in (1, \infty)$ be related by $1/p + 1/q = 1$. Assume $u, u_{x_i} \in L^p_{\text{loc}}(\Omega)$ and $v, v_{x_i} \in L^q_{\text{loc}}(\Omega)$. Then $(uv)_{x_i}$ exists and*

$$(uv)_{x_i} = u_{x_i} v + u\, v_{x_i}.$$

We have the following specialized form of the chain rule.

Proposition 7.1.12 *Assume $f \in C^1(\mathbb{R}, \mathbb{R})$ with f' bounded. Suppose $\Omega \subset \mathbb{R}^d$ is open bounded, and for some $p \in (1, \infty)$, $v \in L^p(\Omega)$ and $v_{x_i} \in L^p(\Omega)$, $1 \le i \le d$, i.e., $v \in W^{1,p}(\Omega)$ using the notation of the Sobolev space $W^{1,p}(\Omega)$ to be introduced in the following section. Then $(f(v))_{x_i} \in L^p(\Omega)$, and $(f(v))_{x_i} = f'(v)\, v_{x_i}$, $1 \le i \le d$.*

Exercise 7.1.1 The "classical" variational lemma, in contrast to the generalized variational lemma of Lemma 7.1.2, is the following (in one dimension): If $v \in C([a, b])$ satisfies

$$\int_a^b v(x)\, \phi(x)\, dx = 0$$

for any $\phi \in C([a, b])$ (or any $\phi \in C^\infty([a, b])$) vanishing in some neighborhood of a and b, then

$$v(x) \equiv 0 \quad \text{for } x \in [a, b].$$

Prove this result.

Exercise 7.1.2 Assume $v \in C([a, b])$ has the property

$$\int_a^b v(x)\, \phi'(x)\, dx = 0$$

for any $\phi \in C^1([a, b])$ with $\phi(a) = \phi(b) = 0$. Show that v is a constant function. *Hint*: Start with

$$\int_a^b [v(x) - c_0]\, \phi'(x)\, dx = 0,$$

where c_0 is the mean value of v over $[a, b]$. Construct ϕ with the required properties such that $\phi'(x) = v(x) - c_0$.

Exercise 7.1.3 Let $\Omega = (0, 1)$, $f(x) = x^{1/2}$. How many weak derivatives does $f(x)$ have? Give formulas of the weak derivatives that exist.

Exercise 7.1.4 Let

$$f(x) = \begin{cases} 1, & -1 < x < 0, \\ ax + b, & 0 \le x < 1. \end{cases}$$

Find a necessary and sufficient condition on a and b for $f(x)$ to be weakly differentiable on $(-1, 1)$. Calculate the weak derivative $f'(x)$ when it exists.

Exercise 7.1.5 Show that if $v \in W^{1,p}(\Omega)$, $1 \le p \le \infty$, then $|v|, v^+, v^- \in W^{1,p}(\Omega)$, and

$$\partial_{x_i} v^+ = \begin{cases} \partial_{x_i} v, & \text{a.e. on } \{\boldsymbol{x} \in \Omega \mid v(\boldsymbol{x}) > 0\}, \\ 0, & \text{a.e. on } \{\boldsymbol{x} \in \Omega \mid v(\boldsymbol{x}) \le 0\}, \end{cases}$$

$$\partial_{x_i} v^- = \begin{cases} 0, & \text{a.e. on } \{\boldsymbol{x} \in \Omega \mid v(\boldsymbol{x}) \ge 0\}, \\ -\partial_{x_i} v, & \text{a.e. on } \{\boldsymbol{x} \in \Omega \mid v(\boldsymbol{x}) < 0\}, \end{cases}$$

where $v^+ = (|v| + v)/2$ is the positive part of v, and $v^- = (|v| - v)/2$ is the negative part.
Hint: To prove the formula for $\partial_{x_i} v^+$, apply Proposition 7.1.12 to the function

$$f_\varepsilon(v) = \begin{cases} (v^2 + \varepsilon^2)^{1/2} - \varepsilon & \text{if } v > 0, \\ 0 & \text{if } v \le 0, \end{cases}$$

where $\varepsilon > 0$, to obtain a formula for $\partial_{x_i} f_\varepsilon(v)$. Then take the limit $\varepsilon \to 0+$.

Exercise 7.1.6 Assume v has the α^{th} weak derivative $w_\alpha = \partial^\alpha u$ and w_α has the β^{th} weak derivative $w_{\alpha+\beta} = \partial^\beta w_\alpha$. Show that $w_{\alpha+\beta}$ is the $(\alpha+\beta)^{th}$ weak derivative of v.

7.2 Sobolev spaces

Some properties of Sobolev spaces require a certain degree of regularity of the boundary $\partial\Omega$ of the domain Ω.

Definition 7.2.1 *Let Ω be open and bounded in $\mathbb{R}^d$, and let V denote a function space on $\mathbb{R}^{d-1}$. We say $\partial\Omega$ is of class V if for each point $\boldsymbol{x}_0 \in \partial\Omega$, there exist an $r > 0$ and a function $g \in V$ such that upon a transformation of the coordinate system if necessary, we have*

$$\Omega \cap B(\boldsymbol{x}_0, r) = \{\boldsymbol{x} \in B(\boldsymbol{x}_0, r) \mid x_d > g(x_1, \ldots, x_{d-1})\}.$$

Here, $B(\boldsymbol{x}_0, r)$ denotes the d-dimensional ball centered at $\boldsymbol{x}_0$ with radius r.

In particular, when V consists of Lipschitz continuous functions, we say Ω is a Lipschitz domain. When V consists of C^k functions, we say Ω is a C^k domain. When V consists of $C^{k,\alpha}$ $(0 < \alpha \le 1)$ functions, we say $\partial\Omega$ is a Hölder boundary of class $C^{k,\alpha}$.

We remark that smooth domains are certainly Lipschitz continuous, and in engineering applications, most domains are Lipschitz continuous (Figures 7.3 and 7.4). Well-known non-Lipschitz domains are the ones with cracks (Figure 7.5).

Since $\partial\Omega$ is a compact set in $\mathbb{R}^d$, we can actually find a finite number of points $\{\boldsymbol{x}_i\}_{i=1}^I$ on the boundary so that for some positive numbers $\{r_i\}_{i=1}^I$ and functions $\{g_i\}_{i=1}^I \subset V$,

$$\Omega \cap B(\boldsymbol{x}_i, r_i) = \{\boldsymbol{x} \in B(\boldsymbol{x}_i, r_i) \mid x_d > g_i(x_1, \ldots, x_{d-1})\}$$

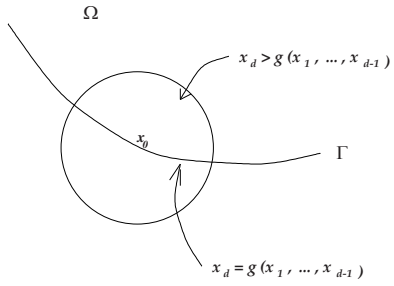

FIGURE 7.2. Smoothness of the boundary

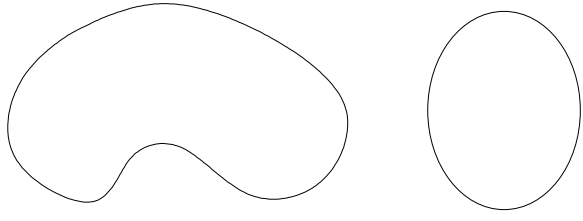

FIGURE 7.3. Smooth domains

upon a transformation of the coordinate system if necessary, and

$$\partial\Omega \subset \bigcup_{i=1}^{I} B(\boldsymbol{x}_i, r_i).$$

7.2.1 Sobolev spaces of integer order

Definition 7.2.2 *Let k be a non-negative integer, $p \in [1, \infty]$. The Sobolev space $W^{k,p}(\Omega)$ is the set of all the functions $v \in L^p(\Omega)$ such that for each multi-index α with $|\alpha| \leq k$, the α^{th} weak derivative $\partial^\alpha v$ exists and $\partial^\alpha v \in L^p(\Omega)$. The norm in the space $W^{k,p}(\Omega)$ is defined as*

$$\|v\|_{W^{k,p}(\Omega)} = \begin{cases} \left[\sum_{|\alpha| \leq k} \|\partial^\alpha v\|^p_{L^p(\Omega)} \right]^{1/p}, & 1 \leq p < \infty, \\ \max_{|\alpha| \leq k} \|\partial^\alpha v\|_{L^\infty(\Omega)}, & p = \infty. \end{cases}$$

When $p = 2$, we write $H^k(\Omega) \equiv W^{k,2}(\Omega)$.

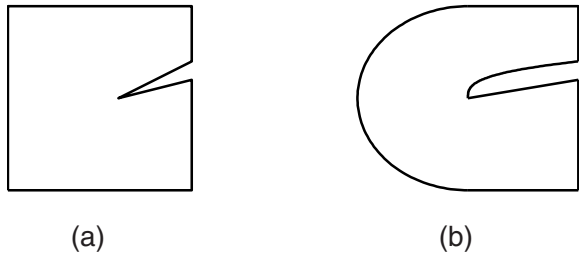

FIGURE 7.4. Lipschitz domains

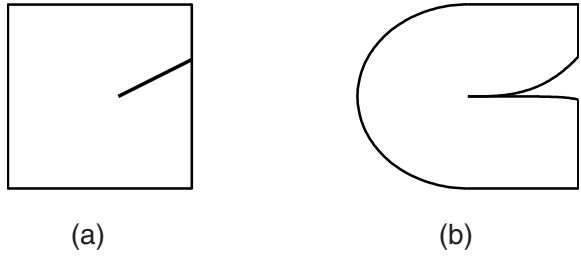

FIGURE 7.5. Crack domains

Usually we replace $\|v\|_{W^{k,p}(\Omega)}$ by the simpler notation $\|v\|_{k,p,\Omega}$, or even $\|v\|_{k,p}$ when no confusion results. When $p = 2$, we similarly use $\|v\|_{k,\Omega}$ or $\|v\|_k$ for the norm $\|v\|_{H^k(\Omega)}$. The standard seminorm over the space $W^{k,p}(\Omega)$ is

$$|v|_{W^{k,p}(\Omega)} = \begin{cases} \left[\displaystyle\sum_{|\alpha|=k} \|\partial^\alpha v\|^p_{L^p(\Omega)} \right]^{1/p}, & 1 \le p < \infty, \\ \displaystyle\max_{|\alpha|=k} \|\partial^\alpha v\|_{L^\infty(\Omega)}, & p = \infty. \end{cases}$$

It is not difficult to see that $W^{k,p}(\Omega)$ is a normed space. Moreover, we have the following result.

Theorem 7.2.3 *The Sobolev space $W^{k,p}(\Omega)$ is a Banach space.*

Proof. Let $\{v_n\}$ be a Cauchy sequence in $W^{k,p}(\Omega)$. Then for any multi-index α with $|\alpha| \le k$, $\{\partial^\alpha v_n\}$ is a Cauchy sequence in $L^p(\Omega)$. Since $L^p(\Omega)$

is complete, there exists a $v_\alpha \in L^p(\Omega)$ such that

$$\partial^\alpha v_n \to v_\alpha \quad \text{in } L^p(\Omega), \text{ as } n \to \infty.$$

Let us show that $v_\alpha = \partial^\alpha v$ where v is the limit of the sequence $\{v_n\}$ in $L^p(\Omega)$. For any $\phi \in C_0^\infty(\Omega)$,

$$\int_\Omega v_n(\boldsymbol{x})\, \partial^\alpha \phi(\boldsymbol{x})\, dx = (-1)^{|\alpha|} \int_\Omega \partial^\alpha v_n(\boldsymbol{x})\, \phi(\boldsymbol{x})\, dx.$$

Letting $n \to \infty$, we obtain

$$\int_\Omega v(\boldsymbol{x})\, \partial^\alpha \phi(\boldsymbol{x})\, dx = (-1)^{|\alpha|} \int_\Omega v_\alpha(\boldsymbol{x})\, \phi(\boldsymbol{x})\, dx \quad \forall\, \phi \in C_0^\infty(\Omega).$$

Therefore, $v_\alpha = \partial^\alpha v$ and $v_n \to v$ in $W^{k,p}(\Omega)$ as $n \to \infty$. □

A simple consequence of the theorem is the following result.

Corollary 7.2.4 *The Sobolev space $H^k(\Omega)$ is a Hilbert space with the inner product*

$$(u, v)_k = \int_\Omega \sum_{|\alpha| \le k} \partial^\alpha u(\boldsymbol{x})\, \partial^\alpha v(\boldsymbol{x})\, dx, \quad u, v \in H^k(\Omega).$$

Like the case for Lebesgue spaces $L^p(\Omega)$, it can be shown that the Sobolev space $W^{k,p}(\Omega)$ is reflexive if and only if $p \in (1, \infty)$.

Let us examine some examples of Sobolev functions.

Example 7.2.5 Assume $\Omega = \{\boldsymbol{x} \in \mathbb{R}^d \mid |\boldsymbol{x}| < 1\}$ is the unit ball, and let $v(\boldsymbol{x}) = |\boldsymbol{x}|^\lambda$, where λ is a real parameter. Let $p \in [1, \infty)$. Notice that

$$\|v\|^p_{L^p(\Omega)} = \int_\Omega |\boldsymbol{x}|^{\lambda\, p} dx = c_d \int_0^1 r^{\lambda\, p + d - 1} dr,$$

where the constant c_d is the surface area of the unit sphere in $\mathbb{R}^d$. So

$$v \in L^p(\Omega) \quad \Longleftrightarrow \quad \lambda > -d/p.$$

It can be verified that the first-order weak derivative v_{x_i} is given by the formula, if v_{x_i} exists,

$$v_{x_i}(\boldsymbol{x}) = \lambda\, |\boldsymbol{x}|^{\lambda - 2} x_i, \quad \boldsymbol{x} \neq \boldsymbol{0},\ 1 \le i \le d.$$

Thus

$$|\nabla v(\boldsymbol{x})| = |\lambda|\, |\boldsymbol{x}|^{\lambda - 1}, \quad \boldsymbol{x} \neq \boldsymbol{0}.$$

Now

$$\||\nabla v|\|^p_{L^p(\Omega)} = |\lambda|^p \int_\Omega |\boldsymbol{x}|^{(\lambda - 1)\, p} dx = c \int_0^1 r^{(\lambda - 1)\, p + d - 1} dr.$$

We see that

$$v \in W^{1,p}(\Omega) \quad \Longleftrightarrow \quad \lambda > 1 - d/p.$$

More generally, for a nonnegative integer k, we have

$$v \in W^{k,p}(\Omega) \quad \Longleftrightarrow \quad \lambda > k - d/p. \tag{7.2.1}$$

Proof of this equivalence is left to the reader (Exercise 7.2.7).

A function with several integrable weak derivatives may have discontinuity. For the function in this example, we have $v \in C(\overline{\Omega})$ if and only if $\lambda \geq 0$. For $d = 2$ and $p \in [1, 2)$, when $0 > \lambda > 1 - 2/p$, we have $v \in W^{1,p}(\Omega)$ and $v \notin C(\Omega)$. □

With the consideration of continuity of functions from Sobolev spaces given in the example, it is now a proper place to make a remark. Since Sobolev spaces are defined through Lebesgue spaces, strictly speaking, an element in a Sobolev space is an equivalence class of measurable functions that are equal a.e. When we say a function from a Sobolev space is continuous, we mean that from the equivalence class of the function, we can find one function which is continuous.

Example 7.2.6 Continuing the discussion at the end of Example 7.2.5, we ask if an $H^1(\Omega)$ function is continuous for $d = 2$. Consider the example

$$v(\boldsymbol{x}) = \log\left(\log\left(\frac{1}{r}\right)\right), \quad \boldsymbol{x} \in \mathbb{R}^2,\ r = |\boldsymbol{x}|$$

with $\Omega = B(0, \beta)$, a circle of radius $\beta < 1$ in the plane. Then

$$\int_\Omega |\nabla v(\boldsymbol{x})|^2 dx = \frac{-2\pi}{\log \beta} < \infty$$

and also easily, $\|v\|_{L^2(\Omega)} < \infty$. Thus $v \in H^1(\Omega)$, but $v(\boldsymbol{x})$ is unbounded as $\boldsymbol{x} \to \boldsymbol{0}$. For conditions ensuring continuity of functions from a Sobolev space, see Theorem 7.3.7. We will see that in the case of $d = 2$, for $p > 2$, a function in $W^{1,p}(\Omega)$ is continuous. □

Example 7.2.7 In the theory of finite elements, we need to analyze the global regularity of a finite element function from its regularity on each element. Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain, partitioned into Lipschitz sub-domains:

$$\overline{\Omega} = \bigcup_{n=1}^{N} \overline{\Omega_n}.$$

Suppose for some non-negative integer k and some real $p \in [1, \infty)$ or $p = \infty$,

$$v|_{\Omega_n} \in W^{k+1,p}(\Omega_n), \quad 1 \leq n \leq N,\ v \in C^k(\overline{\Omega}).$$

Let us show that

$$v \in W^{k+1,p}(\Omega).$$

Evidently, it is enough to prove the result for the case $k = 0$. Thus let $v \in C(\overline{\Omega})$ be such that for each $n = 1, \ldots, N$, $v \in W^{1,p}(\Omega_n)$. For each i, $1 \le i \le d$, we need to show that $\partial_i v$ exists as a weak derivative and belongs to $L^p(\Omega)$. An obvious candidate for $\partial_i v$ is

$$w_i(\boldsymbol{x}) = \begin{cases} \partial_i v(\boldsymbol{x}), & \boldsymbol{x} \in \cup_{n=1}^N \Omega_n, \\ \text{arbitrary}, & \text{otherwise.} \end{cases}$$

Certainly $w_i \in L^p(\Omega)$. So we only need to verify $w_i = \partial_i v$. By definition, we need to prove

$$\int_\Omega w_i \, \phi \, dx = -\int_\Omega v \, \partial_i \phi \, dx \quad \forall \, \phi \in C_0^\infty(\Omega).$$

Denote the unit outward normal vector on $\partial\Omega_n$ by $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^T$ that exists a.e. since Ω_n has a Lipschitz boundary. We have

$$\begin{aligned} \int_\Omega w_i \phi \, dx &= \sum_{n=1}^N \int_{\Omega_n} \partial_i v \, \phi \, dx \\ &= \sum_{n=1}^N \int_{\partial\Omega_n} v|_{\Omega_n} \phi \, \nu_i ds - \sum_{n=1}^N \int_{\Omega_n} v \, \partial_i \phi \, dx \\ &= \sum_{n=1}^N \int_{\partial\Omega_n} v|_{\Omega_n} \, \phi \, \nu_i \, ds - \int_\Omega v \, \partial_i \phi \, dx, \end{aligned}$$

where we apply an integration by parts formula to the integrals on the subdomains Ω_n. Integration by parts formulas are valid for functions from certain Sobolev spaces, to be discussed in Section 7.4. Now the sum of the boundary integrals is zero: Either a portion of $\partial\Omega_n$ is a part of $\partial\Omega$ and $\phi = 0$ along this portion, or the contributions from the adjacent subdomains cancel each other. Thus,

$$\int_\Omega w_i \phi \, dx = -\int_\Omega v \, \partial_i \phi \, dx \quad \forall \, \phi \in C_0^\infty(\Omega).$$

By definition, $w_i = \partial_i v$. □

Example 7.2.8 Continuing the preceding example, let us show that if $v \in W^{k+1,p}(\Omega)$ and $v \in C^k(\overline{\Omega_n})$, $1 \le n \le N$, then $v \in C^k(\overline{\Omega})$. Obviously it is enough to prove the statement for $k = 0$. Let us argue by contradiction. Thus we assume $v \in W^{1,p}(\Omega)$ and $v \in C(\overline{\Omega}_n)$, $1 \le n \le N$, but there are two

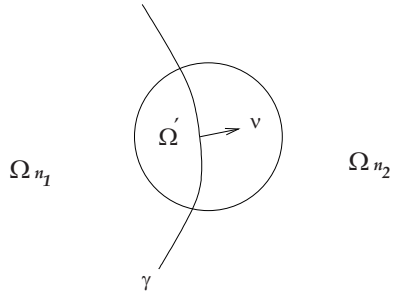

FIGURE 7.6. Two adjacent subdomains

adjacent subdomains Ω_{n_1} and Ω_{n_2} and a nonempty open set $\Omega' \subset \Omega_{n_1} \cup \Omega_{n_2}$ (Figure 7.6) such that

$$v|_{\Omega_{n_1}} > v|_{\Omega_{n_2}} \quad \text{on } \gamma \cap \Omega',$$

where $\gamma = \partial\Omega_{n_1} \cap \partial\Omega_{n_2}$.

By shrinking the set Ω' if necessary, we may assume there is an i between 1 and d, such that $\nu_i > 0$ (or $\nu_i < 0$) on $\gamma \cap \Omega'$. Here ν_i is the i^{th} component of the unit outward normal $\boldsymbol{\nu}$ on γ with respect to Ω_{n_1}. Then we choose $\phi \in C_0^\infty(\Omega') \subset C_0^\infty(\Omega)$ with the property $\phi > 0$ on $\gamma \cap \Omega'$. Now

$$\begin{aligned}
\int_\Omega \partial_i v\, \phi\, dx &= \sum_{l=1}^{2} \int_{\Omega_{n_l}} \partial_i v\, \phi\, dx \\
&= \sum_{l=1}^{2} \int_{\partial\Omega_{n_l}} v|_{\Omega_{n_l}}\, \phi\, \nu_i ds - \sum_{l=1}^{2} \int_{\Omega_{n_l}} v\, \partial_i \phi\, dx \\
&= \int_\gamma (v|_{\Omega_{n_1}} - v|_{\Omega_{n_2}})\, \phi\, \nu_i ds - \int_\Omega v\, \partial_i \phi\, dx.
\end{aligned}$$

By the assumptions, the boundary integral is nonzero. This contradicts the definition of the weak derivative. □

Combining the results from Examples 7.2.7 and 7.2.8, we see that under the assumption $v|_{\Omega_n} \in C^k(\overline{\Omega_n}) \cap W^{k+1,p}(\Omega_n)$, $1 \le n \le N$, we have the conclusion

$$v \in C^k(\overline{\Omega}) \quad \Longleftrightarrow \quad v \in W^{k+1,p}(\Omega).$$

Some important properties of the Sobolev spaces will be discussed in the next section. In general the space $C_0^\infty(\Omega)$ is not dense in $W^{k,p}(\Omega)$. So it is useful to bring in the following definition.

Definition 7.2.9 *Let $W_0^{k,p}(\Omega)$ be the closure of $C_0^\infty(\Omega)$ in $W^{k,p}(\Omega)$. When $p = 2$, we denote $H_0^k(\Omega) \equiv W_0^{k,2}(\Omega)$.*

We interpret $W_0^{k,p}(\Omega)$ to be the space of all the functions v in $W^{k,p}(\Omega)$ with the "property" that

$$\partial^\alpha v(\boldsymbol{x}) = 0 \quad \text{on } \partial\Omega, \ \forall\, \alpha \text{ with } |\alpha| \le k-1.$$

The meaning of this statement is made clear later after the trace theorems are presented.

7.2.2 Sobolev spaces of real order

It is possible to extend the definition of Sobolev spaces with non-negative integer order to any real order. We first introduce Sobolev spaces of positive real order. In this subsection, we assume $p \in [1, \infty)$.

Definition 7.2.10 *Let $s = k + \sigma$ with $k \ge 0$ an integer and $\sigma \in (0, 1)$. Then we define the Sobolev space $W^{s,p}(\Omega)$ to be the set*

$$\left\{ v \in W^{k,p}(\Omega) \;\middle|\; \frac{|\partial^\alpha v(\boldsymbol{x}) - \partial^\alpha v(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|^{\sigma + d/p}} \in L^p(\Omega \times \Omega) \ \forall\, \alpha : |\alpha| = k \right\}$$

with the norm

$$\|v\|_{s,p,\Omega} = \left[\|v\|_{k,p,\Omega}^p + \sum_{|\alpha|=k} \int_{\Omega\times\Omega} \frac{|\partial^\alpha v(\boldsymbol{x}) - \partial^\alpha v(\boldsymbol{y})|^p}{\|\boldsymbol{x} - \boldsymbol{y}\|^{\sigma p + d}}\, dx\, dy \right]^{1/p}.$$

It can be shown that the space $W^{s,p}(\Omega)$ is a Banach space. It is reflexive if and only if $p \in (1, \infty)$. When $p = 2$, $H^s(\Omega) \equiv W^{s,2}(\Omega)$ is a Hilbert space with the inner product

$$(u, v)_{s,\Omega} = (u, v)_{k,\Omega} + \sum_{|\alpha|=k} \int_{\Omega\times\Omega} \frac{[\partial^\alpha u(\boldsymbol{x}) - \partial^\alpha u(\boldsymbol{y})]\,[\partial^\alpha v(\boldsymbol{x}) - \partial^\alpha v(\boldsymbol{y})]}{\|\boldsymbol{x} - \boldsymbol{y}\|^{2\sigma + d}}\, dx\, dy.$$

Most properties of Sobolev spaces of integer order, such as density of smooth functions, extension theorem and Sobolev embedding theorems discussed in the next section, carry over to Sobolev spaces of positive real order introduced here. The introduction of the spaces $W^{s,p}(\Omega)$ in this text serves two purposes: As a preparation for the definition of Sobolev spaces over boundaries and for a more precise statement of Sobolev trace theorems. Therefore, we will not give detailed discussions of the properties of the spaces $W^{s,p}(\Omega)$. An interested reader can consult [137, Chapter 4, Part I].

The space $C_0^\infty(\Omega)$ does not need to be dense in $W^{s,p}(\Omega)$. So we introduce the following definition.

Definition 7.2.11 *Let $s \geq 0$. Then we define $W_0^{s,p}(\Omega)$ to be the closure of the space $C_0^\infty(\Omega)$ in $W^{s,p}(\Omega)$. When $p = 2$, we have a Hilbert space $H_0^s(\Omega) \equiv W_0^{s,2}(\Omega)$.*

With the spaces $W_0^{s,p}(\Omega)$, we can then define Sobolev spaces with negative order.

Definition 7.2.12 *Let $s \geq 0$, either an integer or a non-integer. Let $p \in [1, \infty)$ and denote its conjugate exponent p' defined by the relation $1/p + 1/p' = 1$. Then we define $W^{-s,p'}(\Omega)$ to be the dual space of $W_0^{s,p}(\Omega)$. In particular, $H^{-s}(\Omega) \equiv W^{-s,2}(\Omega)$.*

On several occasions later, we need to use in particular the Sobolev space $H^{-1}(\Omega)$, defined as the dual of $H_0^1(\Omega)$. Thus, any $\ell \in H^{-1}(\Omega)$ is a bounded linear functional on $H_0^1(\Omega)$:

$$|\ell(v)| \leq M \|v\| \quad \forall\, v \in H_0^1(\Omega).$$

The norm of ℓ is

$$\|\ell\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)} \frac{\ell(v)}{\|v\|_{H_0^1(\Omega)}}.$$

Any function $f \in L^2(\Omega)$ naturally induces a bounded linear functional $f \in H^{-1}(\Omega)$ by the relation

$$\langle f, v \rangle = \int_\Omega f\, v\, dx \quad \forall\, v \in H_0^1(\Omega).$$

Sometimes even when $f \in H^{-1}(\Omega) \backslash L^2(\Omega)$, we write $\int_\Omega f\, v\, dx$ for the duality pairing $\langle f, v \rangle$ between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$, although integration in this situation does not make sense.

It can be shown that if $\ell \in H^{-1}(\Omega)$, then there exist $L^2(\Omega)$ functions ℓ_0, $\ldots$, ℓ_d, such that

$$\ell(v) = \int_\Omega \left(\ell_0 v + \sum_{i=1}^d \ell_i v_{x_i} \right) dx \quad \forall\, v \in H_0^1(\Omega).$$

Thus formally (or in the sense of distributions),

$$\ell = \ell_0 - \sum_{i=1}^d \frac{\partial \ell_i}{\partial x_i},$$

i.e., $H^{-1}(\Omega)$ functions can be obtained by differentiating $L^2(\Omega)$ functions.

7.2.3 Sobolev spaces over boundaries

To deal with function spaces over boundaries, we introduce the following Sobolev spaces.

Definition 7.2.13 *Let $k \geq 0$ be an integer, $\alpha \in (0,1]$, $s \in [0, k+\alpha]$ and $p \in [1,\infty)$. Assume a set of local representations of the boundary given by*

$$\partial\Omega \cap B(\boldsymbol{x}_i, r_i) = \{\boldsymbol{x} \in B(\boldsymbol{x}_i, r_i) \mid x_d = g_i(x_1, \ldots, x_{d-1})\}$$

for $i = 1, \ldots, I$, with open $D_i \subset \mathbb{R}^{d-1}$ the domain of g_i; and assume every point of $\partial\Omega$ lies in at least one of these local representations. We assume $g_i \in C^{k,\alpha}(D_i)$ for all i. A decomposition of $\partial\Omega$ into a finite number I of such sub-domains $g_i(D_i)$ is called a "patch system". For $s \leq k+\alpha$, we define the Sobolev space $W^{s,p}(\partial\Omega)$ as follows:

$$W^{s,p}(\partial\Omega) = \left\{ v \in L^2(\partial\Omega) \mid v \circ g_i \in W^{s,p}(D_i),\ i = 1, \ldots, I \right\}.$$

The norm in $W^{s,p}(\partial\Omega)$ is defined by

$$\|v\|_{W^{s,p}(\partial\Omega)} = \max_i \|v \circ g_i\|_{W^{s,p}(D_i)}.$$

Other definitions equivalent to this norm are possible. When $p = 2$, we obtain a Hilbert space $H^s(\partial\Omega) \equiv W^{s,2}(\partial\Omega)$.

Later on, we will mainly use the space $H^{1/2}(\partial\Omega)$, which is further elaborated in Subsection 7.3.4.

Exercise 7.2.1 Show that for non-negative integers k and real $p \in [1,\infty]$, the quantity $\|\cdot\|_{W^{k,p}(\Omega)}$ defines a norm.

Exercise 7.2.2 For the sequence $\{u_n\}$ of zigzag functions illustrated in Figure 7.7, show that as $n \to \infty$,

$$\|u_n\|_{L^2(0,1)} \to 0, \quad \|u_n'\|_{L^2(0,1)} \to \infty.$$

Exercise 7.2.3 Consider the function

$$f(x) = \begin{cases} x^2, & 0 \leq x \leq 1, \\ x^3, & -1 \leq x \leq 0. \end{cases}$$

Determine the largest possible integer k for which $f \in H^k(-1,1)$.

Exercise 7.2.4 Show that $C^k(\overline{\Omega}) \subset W^{k,p}(\Omega)$ for any $p \in [1,\infty]$.

Exercise 7.2.5 Is it true that $C^\infty(\Omega) \subset W^{k,p}(\Omega)$?

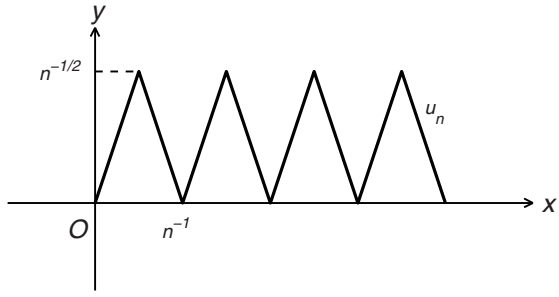

FIGURE 7.7. Zigzag function for Exercise 7.2.2

Exercise 7.2.6 Show that there exists a constant c depending only on k such that

$$\|uv\|_{H^k(\Omega)} \le c\,\|u\|_{C^k(\overline{\Omega})}\|v\|_{H^k(\Omega)} \quad \forall\, u, v \in H^k(\Omega).$$

Hint: Use the formula

$$\partial^\alpha(uv) = \sum_{\beta\le\alpha} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \partial^\beta u \partial^{\alpha-\beta} v.$$

This formula, for both classical derivatives and weak derivatives, can be proved by an inductive argument. See Section 1.4 for the meaning of the quantities involving multi-indices.

Exercise 7.2.7 Prove the relation (7.2.1).

Exercise 7.2.8 Continuing Example 7.2.5, discuss whether it is possible to have a discontinuous function that belongs to $W^{2,p}(\Omega)$.

7.3 Properties

We collect some important properties of the Sobolev spaces in this section. Most properties are stated for Sobolev spaces of nonnegative integer order, although they can be extended to Sobolev spaces of real order. We refer to Sobolev spaces of real order only when it is necessary to do so, e.g., in presentation of trace theorems. Properties of the Sobolev spaces over boundaries are summarized in [137, Chapter 4, Part I].

7.3.1 Approximation by smooth functions

Inequalities involving Sobolev functions are usually proved for smooth functions first, followed by a density argument. A theoretical basis for this technique is density results of smooth functions in Sobolev spaces.

Theorem 7.3.1 *Assume $v \in W^{k,p}(\Omega)$, $1 \le p < \infty$. Then there exists a sequence $\{v_n\} \subset C^\infty(\Omega) \cap W^{k,p}(\Omega)$ such that*

$$\|v_n - v\|_{k,p} \to 0 \quad \text{as } n \to \infty.$$

Note that in this theorem the approximation functions v_n are smooth only in the interior of Ω. To have the smoothness up to the boundary of the approximating sequence, we need to make a smoothness assumption on the boundary of Ω.

Theorem 7.3.2 *Assume Ω is a Lipschitz domain, $v \in W^{k,p}(\Omega)$, $1 \le p < \infty$. Then there exists a sequence $\{v_n\} \subset C^\infty(\overline{\Omega})$ such that*

$$\|v_n - v\|_{k,p} \to 0 \quad \text{as } n \to \infty.$$

Proofs of these density theorems can be found, e.g., in [78].

Since $C^\infty(\overline{\Omega}) \subset C^k(\overline{\Omega}) \subset W^{k,p}(\Omega)$, we see from Theorem 7.3.2 that under the assumption Ω is Lipschitz continuous, the space $W^{k,p}(\Omega)$ is the completion of the space $C^\infty(\overline{\Omega})$ with respect to the norm $\|\cdot\|_{k,p}$.

From the definition of the space $W_0^{k,p}(\Omega)$, we immediately obtain the following density result.

Theorem 7.3.3 *For any $v \in W_0^{k,p}(\Omega)$, there exists a sequence $\{v_n\} \subset C_0^\infty(\Omega)$ such that*

$$\|v_n - v\|_{k,p} \to 0 \quad \text{as } n \to \infty.$$

The definitions of the Sobolev spaces over Ω can be extended in a straightforward fashion to those over the whole space $\mathbb{R}^d$ or other unbounded domains. When $\Omega = \mathbb{R}^d$, smooth functions are dense in Sobolev spaces.

Theorem 7.3.4 *Assume $k \ge 0$, $p \in [1, \infty)$. Then the space $C_0^\infty(\mathbb{R}^d)$ is dense in $W^{k,p}(\mathbb{R}^d)$.*

7.3.2 Extensions

Extension theorems are also useful in proving some relations involving Sobolev functions. A rather general form of extension theorems is the following universal extension theorem, proved in [211, Theorem 5, p. 181].

Theorem 7.3.5 *Assume Ω is an open half-space or a Lipschitz domain in $\mathbb{R}^d$. Then there is an extension operator E such that for any nonnegative integer k and any $p \in [1, \infty]$, E is a linear continuous operator from $W^{k,p}(\Omega)$ to $W^{k,p}(\mathbb{R}^d)$; in other words, for any $v \in W^{k,p}(\Omega)$, we have $Ev \in W^{k,p}(\mathbb{R}^d)$, $Ev = v$ in Ω, Ev is infinitely smooth on $\mathbb{R}^d \backslash \overline{\Omega}$, and*

$$\|Ev\|_{W^{k,p}(\mathbb{R}^d)} \le c \|v\|_{W^{k,p}(\Omega)}$$

for some constant c independent of v.

Notice that in the above theorem, the extension operator E works for all possible values of k and p. In Exercise 7.3.2, we consider a simple extension operator from $W^{k,p}(\mathbb{R}^d_+)$ to $W^{k,p}(\mathbb{R}^d)$, whose definition depends on the value k.

7.3.3 Sobolev embedding theorems

Sobolev embedding theorems are important, e.g. in analyzing the regularity of a weak solution of a boundary value problem.

Definition 7.3.6 *Let V and W be two Banach spaces with $V \subset W$. We say the space V is* continuously embedded *in W and write $V \hookrightarrow W$, if*

$$\|v\|_W \le c\,\|v\|_V \quad \forall\, v \in V. \tag{7.3.1}$$

We say the space V is compactly embedded *in W and write $V \hookrightarrow\hookrightarrow W$, if* (7.3.1) *holds and each bounded sequence in V has a subsequence converging in W.*

If $V \hookrightarrow W$, the functions in V are more smooth than the remaining functions in W. A simple example is $H^1(\Omega) \hookrightarrow L^2(\Omega)$ and $H^1(\Omega) \hookrightarrow\hookrightarrow L^2(\Omega)$. Proofs of most parts of the following two theorems can be found in [78]. The first theorem is on embedding of Sobolev spaces, and the second on compact embedding.

Theorem 7.3.7 *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Then the following statements are valid.*

(a) *If $k < d/p$, then $W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$ for any $q \le p^*$, where p^* is defined by $1/p^* = 1/p - k/d$.*

(b) *If $k = d/p$, then $W^{k,p}(\Omega) \hookrightarrow L^q(\Omega)$ for any $q < \infty$.*

(c) *If $k > d/p$, then*

$$W^{k,p}(\Omega) \hookrightarrow C^{k-[d/p]-1,\beta}(\Omega),$$

where

$$\beta = \begin{cases} [d/p] + 1 - d/p, & \text{if } d/p \neq \text{integer}, \\ \text{any positive number } < 1, & \text{if } d/p = \text{integer}. \end{cases}$$

In the theorem, $[x]$ denotes the integer part of x, i.e. the largest integer less than or equal to x. We remark that in the one-dimensional case, with $\Omega = (a, b)$ a bounded interval, we have

$$W^{k,p}(a, b) \hookrightarrow C[a, b]$$

for any $k \ge 1$, $p \ge 1$.

Theorem 7.3.8 *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Then the following statements are valid.*

(a) *If $k < d/p$, then $W^{k,p}(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega)$ for any $q < p^*$, where p^* is defined by $1/p^* = 1/p - k/d$.*

(b) *If $k = d/p$, then $W^{k,p}(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega)$ for any $q < \infty$.*

(c) *If $k > d/p$, then*

$$W^{k,p}(\Omega) \hookrightarrow\hookrightarrow C^{k-[d/p]-1,\beta}(\Omega),$$

where $\beta \in [0, [d/p] + 1 - d/p)$.

How to remember these results? We take Theorem 7.3.7 as an example. The larger the product kp, the smoother the functions from the space $W^{k,p}(\Omega)$. There is a critical value d (the dimension of the domain Ω) for this product such that if $kp > d$, then a $W^{k,p}(\Omega)$ function is actually continuous (or more precisely, is equal to a continuous function a.e.). When $kp < d$, a $W^{k,p}(\Omega)$ function belongs to $L^{p^*}(\Omega)$ for an exponent p^* larger than p. To determine the exponent p^*, we start from the condition $kp < d$, which is written as $1/p - k/d > 0$. Then $1/p^*$ is defined to be the difference $1/p - k/d$. When $kp > d$, it is usually useful to know if a $W^{k,p}(\Omega)$ function has continuous derivatives up to certain order. We begin with

$$W^{k,p}(\Omega) \hookrightarrow C(\overline{\Omega}) \quad \text{if } k > d/p.$$

Then we apply this embedding result to derivatives of Sobolev functions; it is easy to see that

$$W^{k,p}(\Omega) \hookrightarrow C^l(\overline{\Omega}) \quad \text{if } k - l > d/p.$$

As some concrete examples, for a two-dimensional Lipschitz domain Ω, $H^1(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega)$ $\forall\, 1 \le q < \infty$ and $H^2(\Omega) \hookrightarrow\hookrightarrow C(\overline{\Omega})$. So in particular, a sequence bounded in $H^1(\Omega)$ has a subsequence that converges in $L^2(\Omega)$, and a sequence bounded in $H^2(\Omega)$ has a subsequence that converges in $C(\overline{\Omega})$. For a three-dimensional Lipschitz domain Ω, $H^1(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega)$ $\forall\, 1 \le q < 6$, $H^1(\Omega) \hookrightarrow L^6(\Omega)$, and $H^2(\Omega) \hookrightarrow\hookrightarrow C(\overline{\Omega})$.

A direct consequence of Theorem 7.3.8 is the following compact embedding result.

Theorem 7.3.9 *Let k and l be non-negative integers, $k > l$, and $p \in [1, \infty]$. Let $\Omega \subset \mathbb{R}^d$ be a non-empty open bounded Lipschitz domain. Then $W^{k,p}(\Omega) \hookrightarrow\hookrightarrow W^{l,p}(\Omega)$.*

7.3.4 Traces

Sobolev spaces are defined through $L^p(\Omega)$ spaces. Hence Sobolev functions are uniquely defined only a.e. in Ω. Now that the boundary Γ has measure zero in $\mathbb{R}^d$, it seems the boundary value of a Sobolev function is not well-defined. Nevertheless it is possible to define the trace of a Sobolev function on the boundary in such a way that for a Sobolev function that is continuous up to the boundary, its trace coincides with its boundary value.

Theorem 7.3.10 *Assume Ω is a Lipschitz domain in $\mathbb{R}^d$, $1 \le p < \infty$. Then there exists a continuous linear operator $\gamma : W^{1,p}(\Omega) \to L^p(\Gamma)$ with the following properties:*

(a) $\gamma v = v|_\Gamma$ *if* $v \in W^{1,p}(\Omega) \cap C(\overline{\Omega})$.

(b) *For some constant* $c > 0$, $\|\gamma v\|_{L^p(\Gamma)} \le c\,\|v\|_{W^{1,p}(\Omega)}\ \forall\, v \in W^{1,p}(\Omega)$.

(c) *The mapping $\gamma : W^{1,p}(\Omega) \to L^p(\Gamma)$ is compact; i.e., for any bounded sequence $\{v_n\}$ in $W^{1,p}(\Omega)$, there is a subsequence $\{v_{n'}\} \subset \{v_n\}$ such that $\{\gamma v_{n'}\}$ is convergent in $L^p(\Gamma)$.*

The operator γ is called the *trace operator*, and γv can be called the *generalized boundary value* of v. The trace operator is neither an injection nor a surjection from $W^{1,p}(\Omega)$ to $L^p(\Gamma)$. The range $\gamma(W^{1,p}(\Omega))$ is a space smaller than $L^p(\Gamma)$, namely $W^{1-1/p,p}(\Gamma)$, a positive order Sobolev space over the boundary. Usually we use the same symbol v for the trace of $v \in H^1(\Omega)$.

As a consequence of Theorem 7.3.10, if $v_n \to v$ in $H^1(\Omega)$, then $v_n \to v$ in $L^2(\Gamma)$.

When we discuss weak formulations of boundary value problems later in this book, we need to use traces of the $H^1(\Omega)$ functions. These traces form the space $H^{1/2}(\Gamma)$; in other words,

$$H^{1/2}(\Gamma) = \gamma(H^1(\Omega)).$$

Correspondingly, we can use the following as the norm for $H^{1/2}(\Gamma)$:

$$\|g\|_{H^{1/2}(\Gamma)} = \inf_{\substack{v \in H^1(\Omega) \\ \gamma v = g}} \|v\|_{H^1(\Omega)}. \tag{7.3.2}$$

In studying boundary value problems, necessarily we need to be able to impose essential boundary conditions properly in formulations. For second-order boundary value problems, essential boundary conditions involve only function values on the boundary, so Theorem 7.3.10 is sufficient for the purpose. For higher-order boundary value problems, we need to use the traces of partial derivatives on the boundary. For example, for fourth-order boundary value problems, any boundary conditions involving derivatives of order at most one are treated as essential boundary conditions. Since

a tangential derivative of a function on the boundary can be obtained by taking a differentiation of the boundary value of the function, we only need to use traces of a function and its normal derivative.

Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^T$ denote the outward unit normal to the boundary Γ of Ω. Recall that if $v \in C^1(\overline{\Omega})$, then its classical normal derivative on the boundary is

$$\frac{\partial v}{\partial \nu} = \sum_{i=1}^{d} \frac{\partial v}{\partial x_i} \nu_i.$$

The following theorem ([98]) states the fact that for a function from certain Sobolev spaces, it is possible to define a *generalized normal derivative* which is an extension of the classical normal derivative.

Theorem 7.3.11 *Assume Ω is a bounded open set with a $C^{k,1}$ boundary Γ. Assume $1 \le p \le \infty$, $s - 1/p > 1$ is not an integer, and $k \ge s - 1$. Then there exist unique bounded linear and surjective mappings $\gamma_0 : W^{s,p}(\Omega) \to W^{s-1/p,p}(\Gamma)$ and $\gamma_1 : W^{s,p}(\Omega) \to W^{s-1-1/p,p}(\Gamma)$ such that $\gamma_0 v = v|_\Gamma$ and $\gamma_1 v = (\partial v/\partial \nu)|_\Gamma$ when $v \in W^{s,p}(\Omega) \cap C^1(\overline{\Omega})$.*

7.3.5 Equivalent norms

In the study of weak formulations of boundary value problems, it is convenient to use equivalent norms over Sobolev spaces or Sobolev subspaces. There are some powerful general results, called norm equivalence theorems, for the purpose of generating various equivalent norms on Sobolev spaces. Before stating the norm equivalence results, we recall the seminorm defined by

$$|v|_{k,p,\Omega} = \left(\int_\Omega \sum_{|\alpha|=k} |\partial^\alpha v|^p dx \right)^{1/p}$$

over the space $W^{k,p}(\Omega)$ for $p < \infty$. It can be shown that if Ω is connected and $|v|_{k,p,\Omega} = 0$, then $v \in \mathbb{P}_{k-1}(\Omega)$.

Theorem 7.3.12 *Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, $k \ge 1$, $1 \le p < \infty$. Assume $f_j : W^{k,p}(\Omega) \to \mathbb{R}$, $1 \le j \le J$, are seminorms on $W^{k,p}(\Omega)$ satisfying two conditions:*

(H1) $0 \le f_j(v) \le c \|v\|_{k,p,\Omega} \quad \forall\, v \in W^{k,p}(\Omega)$, $1 \le j \le J$.

(H2) *If $v \in \mathbb{P}_{k-1}(\Omega)$ and $f_j(v) = 0$, $1 \le j \le J$, then $v = 0$.*

Then the quantity

$$\|v\| = |v|_{k,p,\Omega} + \sum_{j=1}^{J} f_j(v) \tag{7.3.3}$$

or

$$\|v\| = \left[|v|_{k,p,\Omega}^p + \sum_{j=1}^{J} f_j(v)^p \right]^{1/p} \tag{7.3.4}$$

defines a norm on $W^{k,p}(\Omega)$, which is equivalent to the norm $\|v\|_{k,p,\Omega}$.

Proof. We prove that the quantity (7.3.3) is a norm on $W^{k,p}(\Omega)$ equivalent to the norm $\|v\|_{k,p,\Omega}$. The statement on the quantity (7.3.4) can be proved similarly or by noting the equivalence between the two quantities (7.3.3) and (7.3.4).

By the condition (H1), we see that for some constant $c > 0$,

$$\|v\| \le c\,\|v\|_{k,p,\Omega} \quad \forall\, v \in W^{k,p}(\Omega).$$

So we only need to show that there is another constant $c > 0$ such that

$$\|v\|_{k,p,\Omega} \le c\,\|v\| \quad \forall\, v \in W^{k,p}(\Omega).$$

We argue by contradiction. Suppose this inequality is false. Then we can find a sequence $\{v_n\} \subset W^{k,p}(\Omega)$ with the properties

$$\|v_n\|_{k,p,\Omega} = 1, \tag{7.3.5}$$

$$\|v_n\| \le 1/n \tag{7.3.6}$$

for $n = 1, 2, \ldots$. From (7.3.6), we see that as $n \to \infty$,

$$|v_n|_{k,p,\Omega} \to 0 \tag{7.3.7}$$

and

$$f_j(v_n) \to 0, \quad 1 \le j \le J. \tag{7.3.8}$$

Since $\{v_n\}$ is a bounded sequence in $W^{k,p}(\Omega)$ from the property (7.3.5), and since

$$W^{k,p}(\Omega) \hookrightarrow\hookrightarrow W^{k-1,p}(\Omega),$$

there is a subsequence of the sequence $\{v_n\}$, still denoted as $\{v_n\}$, and a function $v \in W^{k-1,p}(\Omega)$ such that

$$v_n \to v \quad \text{in } W^{k-1,p}(\Omega), \text{ as } n \to \infty. \tag{7.3.9}$$

This property and (7.3.7), together with the uniqueness of a limit, imply that

$$v_n \to v \quad \text{in } W^{k,p}(\Omega), \text{ as } n \to \infty$$

and

$$|v|_{k,p,\Omega} = \lim_{n\to\infty} |v_n|_{k,p,\Omega} = 0.$$

We then conclude that $v \in \mathbb{P}_{k-1}(\Omega)$. On the other hand, from the continuity of the functionals $\{f_j\}_{1\le j\le J}$ and (7.3.8), we find that

$$f_j(v) = \lim_{n\to\infty} f_j(v_n) = 0, \quad 1 \le j \le J.$$

Using the condition (H2), we see that $v = 0$, which contradicts the relation that

$$\|v\|_{k,p,\Omega} = \lim_{n\to\infty} \|v_n\|_{k,p,\Omega} = 1.$$

The proof of the result is now completed. □

Notice that in Theorem 7.3.12, we need to assume Ω to be connected. This assumption is used to guarantee that from $|v|_{k,p,\Omega} = 0$ we can conclude that v is a (global) polynomial of degree less than or equal to $k-1$. The above proof of Theorem 7.3.12 can be easily modified to yield the next result.

Theorem 7.3.13 *Let Ω be an open, bounded set in $\mathbb{R}^d$ with a Lipschitz boundary, $k \ge 1$, $1 \le p < \infty$. Assume $f_j : W^{k,p}(\Omega) \to \mathbb{R}$, $1 \le j \le J$, are seminorms on $W^{k,p}(\Omega)$ satisfying two conditions:*

(H1) $0 \le f_j(v) \le c\,\|v\|_{k,p,\Omega} \quad \forall\, v \in W^{k,p}(\Omega),\ 1 \le j \le J.$

(H2)′ *If $|v|_{k,p,\Omega} = 0$ and $f_j(v) = 0$, $1 \le j \le J$, then $v = 0$.*

Then the quantities (7.3.3) *and* (7.3.4) *are norms on $W^{k,p}(\Omega)$, equivalent to the norm $\|v\|_{k,p,\Omega}$.*

We may also state the norm-equivalence result for the case where Ω is a union of separated open connected sets.

Theorem 7.3.14 *Let Ω be an open, bounded set in $\mathbb{R}^d$, $\Omega = \cup_{\lambda\in\Lambda}\Omega_\lambda$ with each Ω_λ having a Lipschitz boundary, $\Omega_\lambda \cap \Omega_\mu = \emptyset$ for $\lambda \ne \mu$. Let $k \ge 1$, $1 \le p < \infty$. Assume $f_j : W^{k,p}(\Omega) \to \mathbb{R}$, $1 \le j \le J$, are seminorms on $W^{k,p}(\Omega)$ satisfying two conditions:*

(H1) $0 \le f_j(v) \le c\,\|v\|_{k,p,\Omega}\ \forall\, v \in W^{k,p}(\Omega),\ 1 \le j \le J.$

(H2) *If $v|_{\Omega_\lambda} \in \mathbb{P}_{k-1}(\Omega_\lambda)$, $\forall\,\lambda \in \Lambda$, and $f_j(v) = 0$, $1 \le j \le J$, then $v = 0$.*

Then the quantities (7.3.3) *and* (7.3.4) *are norms on $W^{k,p}(\Omega)$, equivalent to the norm $\|v\|_{k,p,\Omega}$.*

Many useful inequalities can be derived as consequences of the previous theorems. We present some examples below.

Example 7.3.15 Assume Ω is an open, bounded set in $\mathbb{R}^d$ with a Lipschitz boundary. Let us apply Theorem 7.3.13 with $k = 1$, $p = 2$, $J = 1$ and

$$f_1(v) = \int_\Gamma |v|\, ds.$$

We can then conclude that there exists a constant $c > 0$, depending only on Ω such that

$$\|v\|_{1,\Omega} \leq c \left[|v|_{1,\Omega} + \|v\|_{L^1(\Gamma)} \right] \quad \forall\, v \in H^1(\Omega).$$

Therefore, the Poincaré-Friedrichs inequality holds:

$$\|v\|_{1,\Omega} \leq c\, |v|_{1,\Omega} \quad \forall\, v \in H^1_0(\Omega). \tag{7.3.10}$$

From this inequality it follows that the seminorm $|\cdot|_{1,\Omega}$ is a norm on $H^1_0(\Omega)$, equivalent to the usual $H^1(\Omega)$-norm. □

An extension of the inequality (7.3.10) is discussed in Exercise 7.3.5.

Example 7.3.16 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Assume Γ_0 is an open, non-empty subset of the boundary Γ. Then there is a constant $c > 0$, depending only on Ω, such that

$$\|v\|_{1,\Omega} \leq c \left[|v|_{1,\Omega} + \|v\|_{L^1(\Gamma_0)} \right] \quad \forall\, v \in H^1(\Omega).$$

This inequality can be derived by applying Theorem 7.3.12 with $k = 1$, $p = 2$, $J = 1$ and

$$f_1(v) = \int_{\Gamma_0} |v|\, ds.$$

Therefore,

$$\|v\|_{1,\Omega} \leq c\, |v|_{1,\Omega} \quad \forall\, v \in H^1_{\Gamma_0}(\Omega),$$

where

$$H^1_{\Gamma_0}(\Omega) = \{ v \in H^1(\Omega) \mid v = 0 \text{ a.e. on } \Gamma_0 \}$$

is a subspace of $H^1(\Omega)$. □

Some other useful inequalities, however, cannot be derived from the norm equivalence theorem. One example is

$$\|v\|_{2,\Omega} \leq c\, |\Delta v|_{0,\Omega} \quad \forall\, v \in H^2(\Omega) \cap H^1_0(\Omega) \tag{7.3.11}$$

which is valid if Ω is smooth or is convex. This result is proved by using a regularity estimate for the (weak) solution of the boundary value problem (see [78])

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Another example is Korn's inequality useful in theoretical mechanics. Let Ω be a Lipschitz domain in $\mathbb{R}^3$. Given a function $\boldsymbol{u} \in [H^1(\Omega)]^3$, the linearized strain tensor is defined by

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2} \left[\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T \right];$$

in component form,

$$\varepsilon_{ij}(\boldsymbol{u}) = \frac{1}{2}\left(\partial_{x_i} u_j + \partial_{x_j} u_i\right), \quad 1 \le i, j \le 3.$$

Let Γ_0 be a measurable subset of $\partial\Omega$ with $\text{meas}\,(\Gamma_0) > 0$, and define

$$[H^1_{\Gamma_0}(\Omega)]^3 = \left\{\boldsymbol{v} \in [H^1(\Omega)]^3 \mid \boldsymbol{v} = \boldsymbol{0} \text{ a.e. on } \Gamma_0\right\}.$$

Korn's inequality states that there exists a constant $c > 0$ depending only on Ω such that

$$\|\boldsymbol{v}\|^2_{[H^1(\Omega)]^3} \le c \int_\Omega |\boldsymbol{\varepsilon}(\boldsymbol{v})|^2 dx \quad \forall\, \boldsymbol{v} \in [H^1_{\Gamma_0}(\Omega)]^3. \tag{7.3.12}$$

A proof of Korn's inequality can be found in [143] or [177].

7.3.6 A Sobolev quotient space

Later in error analysis for the finite element method, we need an inequality involving the norm of the Sobolev quotient space

$$\begin{aligned} V &= W^{k+1,p}(\Omega)/\mathbb{P}_k(\Omega) \\ &= \{[v] \mid [v] = \{v + q \mid q \in \mathbb{P}_k(\Omega)\},\ v \in W^{k+1,p}(\Omega)\}. \end{aligned}$$

Here $k \ge 0$ is an integer. Any element $[v]$ of the space V is an equivalence class, the difference between any two elements in the equivalence class being a polynomial in the space $\mathbb{P}_k(\Omega)$. Any $v \in [v]$ is called a representative element of $[v]$. The quotient norm in the space V is defined to be

$$\|[v]\|_V = \inf_{q \in \mathbb{P}_k(\Omega)} \|v + q\|_{k+1,p,\Omega}.$$

Theorem 7.3.17 *Assume $1 \le p < \infty$. Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Then the quantity $|v|_{k+1,p,\Omega}$, $\forall\, v \in [v]$, is a norm on V, equivalent to the quotient norm $\|[v]\|_V$.*

Proof. Obviously, for any $[v] \in V$ and any $v \in [v]$,

$$\|[v]\|_V = \inf_{q \in \mathbb{P}_k(\Omega)} \|v + q\|_{k+1,p,\Omega} \ge |v|_{k+1,p,\Omega}.$$

Thus we only need to prove that there is a constant c, depending only on Ω, such that

$$\inf_{q \in \mathbb{P}_k(\Omega)} \|v + q\|_{k+1,p,\Omega} \le c\,|v|_{k+1,p,\Omega} \quad \forall\, v \in W^{k+1,p}(\Omega). \tag{7.3.13}$$

Denote $N = \dim(\mathbb{P}_k(\Omega))$. Define N independent linear continuous functionals on $\mathbb{P}_k(\Omega)$, the continuity being with respect to the norm of $W^{k+1,p}(\Omega)$.

By the Hahn-Banach theorem, we can extend these functionals to linear continuous functionals over the space $W^{k+1,p}(\Omega)$, denoted by $f_i(\cdot)$, $1 \le i \le N$, such that for $q \in \mathbb{P}_k(\Omega)$, $f_i(q) = 0$, $1 \le i \le N$, if and only if $q = 0$. Then $|f_i(\cdot)|$, $1 \le i \le N$, are seminorms on $W^{k+1,p}(\Omega)$ satisfying the assumptions of Theorem 7.3.12. Applying Theorem 7.3.12, we have

$$\|v\|_{k+1,p,\Omega} \le c \left[|v|_{k+1,p,\Omega} + \sum_{i=1}^{N} |f_i(v)| \right] \quad \forall\, v \in W^{k+1,p}(\Omega).$$

Since f_i, $1 \le i \le N$, are linearly independent on $\mathbb{P}_k(\Omega)$, for any fixed $v \in W^{k+1,p}(\Omega)$, there exists $q \in \mathbb{P}_k(\Omega)$ such that $f_i(v+q) = 0$, $1 \le i \le N$. Thus,

$$\|v+q\|_{k+1,p,\Omega} \le c \left[|v+q|_{k+1,p,\Omega} + \sum_{i=1}^{N} |f_i(v+q)| \right] = c\,|v|_{k+1,p,\Omega},$$

and hence (7.3.13) holds.

It is possible to prove (7.3.13) without using the Hahn-Banach theorem. For this, we apply Theorem 7.3.12 to obtain the inequality

$$\|v\|_{k+1,p,\Omega} \le c \left[|v|_{k+1,p,\Omega} + \sum_{|\alpha| \le k} \left| \int_\Omega \partial^\alpha v(\boldsymbol{x})\, dx \right| \right] \quad \forall\, v \in W^{k+1,p}(\Omega).$$

Replacing v by $v+q$ and noting that $\partial^\alpha q = 0$ for $|\alpha| = k+1$, we have

$$\|v+q\|_{k+1,p,\Omega} \le c \left[|v|_{k+1,p,\Omega} + \sum_{|\alpha| \le k} \left| \int_\Omega \partial^\alpha (v+q)\, dx \right| \right] \qquad (7.3.14)$$
$$\forall\, v \in W^{k+1,p}(\Omega),\ q \in \mathbb{P}_k(\Omega).$$

Now construct a polynomial $\bar{q} \in \mathbb{P}_k(\Omega)$ satisfying

$$\int_\Omega \partial^\alpha (v + \bar{q})\, dx = 0 \quad \text{for } |\alpha| \le k. \qquad (7.3.15)$$

This can always be done: Set $|\alpha| = k$, then $\partial^\alpha \bar{p}$ equals $\alpha_1! \cdots \alpha_d!$ times the coefficient of $\boldsymbol{x}^\alpha \equiv x_1^{\alpha_1} \cdots x_d^{\alpha_d}$, so the coefficient can be computed by using (7.3.15). Having found all the coefficients for terms of degree k, we set $|\alpha| = k-1$, and use (7.3.15) to compute all the coefficients for terms of degree $k-1$. Proceeding in this way, we obtain the polynomial $\bar{q}$ satisfying the condition (7.3.15) for the given function v.

With $q = \bar{q}$ in (7.3.14), we have

$$\inf_{q \in \mathbb{P}_k(\Omega)} \|v+q\|_{k+1,p,\Omega} \le \|v+\bar{q}\|_{k+1,p,\Omega} \le c\,|v|_{k+1,p,\Omega},$$

from which (7.3.13) follows. □

Corollary 7.3.18 *For any Lipschitz domain $\Omega \subset \mathbb{R}^d$, there is a constant c, depending only on Ω, such that*

$$\inf_{p\in\mathbb{P}_k(\Omega)} \|v+p\|_{k+1,\Omega} \le c\,|v|_{k+1,\Omega} \quad \forall\, v \in H^{k+1}(\Omega). \tag{7.3.16}$$

Exercise 7.3.1 Assume Ω is a Lipschitz domain, $v \in W^{k,p}(\Omega)$, $k \ge 0$ integer, $1 \le p < \infty$. Show that there exists a sequence of polynomials $\{v_n\}_{n\ge1}$ such that

$$\|v - v_n\|_{k,p} \to 0 \quad \text{as } n \to \infty.$$

Hint: Apply Theorem 7.3.2 and recall Exercise 3.1.8.

Exercise 7.3.2 It is possible to construct a simple extension operator when the domain is a half-space, say $\mathbb{R}^d_+ = \{\boldsymbol{x} \in \mathbb{R}^d \mid x_d \ge 0\}$. Let $k \ge 1$ be an integer, $p \in [1,\infty]$. For any $v \in W^{k,p}(\mathbb{R}^d_+)$, we define

$$Ev(\boldsymbol{x}) = \begin{cases} v(\boldsymbol{x}), & \boldsymbol{x} \in \mathbb{R}^d_+, \\ \sum_{j=0}^{k-1} c_j v(x_1,\ldots,x_{d-1},-2^j x_d), & \boldsymbol{x} \in \mathbb{R}^d \backslash \mathbb{R}^d_+, \end{cases}$$

where the coefficients $c_0,\ldots,c_{k-1}$ are determined from the system

$$\sum_{j=0}^{k-1} c_j (-2^j)^i = 1, \quad i = 0,1,\ldots,k-1.$$

Show that $Ev \in W^{k,p}(\mathbb{R}^d)$, and E is a continuous operator from $W^{k,p}(\mathbb{R}^d_+)$ to $W^{k,p}(\mathbb{R}^d)$.

Exercise 7.3.3 In general, an embedding result (Theorems 7.3.7, 7.3.8) is not easy to prove. Nevertheless, it is usually not difficult to prove an embedding result for one-dimensional domains. Let $-\infty < a < b < \infty$, $p > 1$, and let q be the conjugate of p defined by the relation $1/p + 1/q = 1$. Prove the embedding result $W^{1,p}(a,b) \hookrightarrow C^{0,1/q}(a,b)$ with the following steps.

First, let $v \in C^1[a,b]$. By the Mean Value Theorem in Calculus, there exists a $\xi \in [a,b]$ such that

$$\int_a^b v(x)\,dx = (b-a)\,v(\xi).$$

Then we can write

$$v(x) = \frac{1}{b-a}\int_a^b v(s)\,ds + \int_\xi^x v'(s)\,ds,$$

from which it is easy to find

$$|v(x)| \le c\,\|v\|_{W^{1,p}(a,b)} \quad \forall\, x \in [a,b].$$

Hence,

$$\|v\|_{C[a,b]} \le c\,\|v\|_{W^{1,p}(a,b)}.$$

Furthermore, for $x \neq y$,

$$|v(x) - v(y)| = \left| \int_y^x v'(s)\, ds \right| \leq |x-y|^{1/q} \left[\int_a^b |v'(s)|^p ds \right]^{1/p}.$$

Therefore,

$$\|v\|_{C^{0,1/q}(a,b)} \leq c\, \|v\|_{W^{1,p}(a,b)} \quad \forall\, v \in C^1[a,b].$$

Second, for any $v \in W^{1,p}(a,b)$, using the density of $C^1[a,b]$ in $W^{1,p}(a,b)$, we can find a sequence $\{v_n\} \subset C^1[a,b]$ such that

$$\|v_n - v\|_{W^{1,p}(a,b)} \to 0 \quad \text{as } n \to \infty.$$

Apply the inequality proved in the first step,

$$\|v_m - v_n\|_{C^{0,1/q}(a,b)} \leq c\, \|v_m - v_n\|_{W^{1,p}(a,b)} \to 0 \quad \text{as } m, n \to \infty.$$

So $\{v_n\}$ is a Cauchy sequence in $C^{0,1/q}(a,b)$. Since the space $C^{0,1/q}(a,b)$ is complete, the sequence $\{v_n\}$ converges to some $\tilde{v}$ in $C^{0,1/q}(a,b)$. We also have $v_n \to v$ a.e., at least for a subsequence of $\{v_n\}$. By the uniqueness of a limit, we conclude $\tilde{v} = v$.

Exercise 7.3.4 Prove Theorem 7.3.9 by applying Theorem 7.3.8.

Exercise 7.3.5 Assume $\Omega \subset \mathbb{R}^d$ is bounded along the direction of one axis, $p \in [1, \infty)$. Prove the Poincaré inequality

$$\|v\|_{L^p(\Omega)} \leq c\, \|\nabla v\|_{L^p(\Omega)} \quad \forall\, v \in W_0^{1,p}(\Omega).$$

Hint: Due to the density of $C_0^\infty(\Omega)$ in $W_0^{1,p}(\Omega)$, it is sufficient to prove the inequality for $v \in C_0^\infty(\Omega)$. With a change of the coordinate system if necessary, we may assume for some $\ell > 0$ that $\Omega \subset \mathbb{R}^{d-1} \times (-\ell, \ell)$. Extend v by zero outside Ω and write

$$v(\boldsymbol{x}) = \int_{-\ell}^{x_d} \frac{\partial v}{\partial x_d}(x_1, \ldots, x_{d-1}, z)\, dz.$$

Exercise 7.3.6 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, $1 \leq p < \infty$. Then there exists a constant $c > 0$ such that

$$\|v - m_\Omega(v)\|_{L^p(\Omega)} \leq c\, \|\nabla v\|_{L^p(\Omega)} \quad \forall\, v \in W^{1,p}(\Omega),$$

where

$$m_\Omega(v) = \frac{1}{\text{meas}(\Omega)} \int_\Omega v(\boldsymbol{x})\, dx$$

is the mean value of v over Ω. This result is termed Poincaré-Wirtinger inequality.

Exercise 7.3.7 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain with boundary Γ. Assume $\Gamma_0 \subset \Gamma$ is such that $\text{meas}(\Gamma_0) > 0$. Define

$$H^2_{\Gamma_0}(\Omega) = \{v \in H^2(\Omega) \mid v = \partial v/\partial \nu = 0 \text{ a.e. on } \Gamma_0\}.$$

Prove the following inequality

$$\|v\|_{2,\Omega} \le c\,|v|_{2,\Omega} \quad \forall\, v \in H^2_{\Gamma_0}(\Omega).$$

This result implies that under the stated assumptions, $|v|_{2,\Omega}$ is a norm on $H^2_{\Gamma_0}(\Omega)$, which is equivalent to the norm $\|v\|_{2,\Omega}$.

Exercise 7.3.8 Define a subspace of $H^2(a,b)$:

$$V = \left\{ v \in H^2(a,b) \mid v(a) = 0,\, v'(b) + gv(b) = 0 \right\},$$

where $g \in \mathbb{R}$. Discuss whether $|v|_{H^2(a,b)}$ defines a norm over V which is equivalent to $\|v\|_{H^2(a,b)}$.

Exercise 7.3.9 Apply the norm equivalence theorems to derive the following inequalities, stating preciously assumptions on Ω:

$$\|v\|_{1,p,\Omega} \le c \left(|v|_{1,p,\Omega} + \Big| \int_{\Omega_0} v\,dx \Big| \right) \quad \forall\, v \in W^{1,p}(\Omega),$$
$$\text{if } \Omega_0 \subset \Omega,\ \text{meas}(\Omega_0) > 0;$$
$$\|v\|_{1,p,\Omega} \le c \left[|v|_{1,p,\Omega} + \|v\|_{L^p(\Gamma)} \right] \quad \forall\, v \in W^{1,p}(\Omega),$$
$$\text{if } \Gamma \subset \Gamma,\ \text{meas}_{d-1}(\Gamma) > 0.$$
$$\|v\|_{1,p,\Omega} \le c\,|v|_{1,p,\Omega} \quad \forall\, v \in W^{1,p}_0(\Omega).$$

Can you think of some more inequalities of the above kind?

Exercise 7.3.10 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, $\Gamma_0 \subset \Gamma$ a non-trivial part of the boundary, i.e., $\text{meas}_{d-1}(\Gamma_0) > 0$. Assume $1 \le p < \infty$. Show that there exist two positive constants c_1 and c_2 such that if $u, v \in W^{1,p}(\Omega)$ with $u = v$ on Γ_0, then

$$|u|_{1,p,\Omega} \ge c_1 \|u\|_{1,p,\Omega} - c_2 \|v\|_{1,p,\Omega}.$$

Exercise 7.3.11 In some applications, it is important to find or estimate the best constant in a Sobolev inequality. For example, let Ω be a Lipschitz domain, and let Γ_1 and Γ_2 be two disjoint, nonempty open subsets of the boundary Γ. Then there is a Sobolev inequality

$$\|v\|_{L^2(\Gamma_1)} \le c\,\|\nabla v\|_{L^2(\Omega)} \quad \forall\, v \in H^1_{\Gamma_2}(\Omega).$$

By the best constant c_0 of the inequality, we mean that c_0 is the smallest constant such that the inequality holds. The best constant c_0 can be characterized by the expression

$$c_0 = \sup\{ \|v\|_{L^2(\Gamma_1)} / \|\nabla v\|_{L^2(\Omega)} \mid v \in H^1_{\Gamma_2}(\Omega) \}.$$

Show that $c_0 = 1/\sqrt{\lambda_1}$ where $\lambda_1 > 0$ is the smallest eigenvalue of the eigenvalue problem

$$u \in H^1_{\Gamma_2}(\Omega), \quad \int_\Omega \nabla u \cdot \nabla v\,dx = \lambda \int_{\Gamma_1} u\,v\,ds \quad \forall\, v \in H^1_{\Gamma_2}(\Omega).$$

Hint: To do the last part of the exercise, first show the existence of $0 \neq u \in H^1_{\Gamma_2}(\Omega)$ such that

$$c_0 = \|u\|_{L^2(\Gamma_1)} / \|\nabla u\|_{L^2(\Omega)}.$$

This is shown through a sequence $\{v_n\} \subset H^1_{\Gamma_2}(\Omega)$ with $\|\nabla v_n\|_{L^2(\Omega)} = 1$ and $\|v_n\|_{L^2(\Gamma_1)} \to c_0$. A subsequence of $\{v_n\}$ converges weakly in $H^1(\Omega)$ and strongly in $L^2(\Gamma_1)$ to $u \in H^1_{\Gamma_2}(\Omega)$. Then, with this u, for any $v \in H^1_{\Gamma_2}(\Omega)$, the function

$$f(t) = \|u + tv\|^2_{L^2(\Gamma_1)} / \|\nabla(u + tv)\|^2_{L^2(\Omega)}$$

has a maximum at $t = 0$; this leads to the conclusion that u is an eigenfunction corresponding to the eigenvalue λ_1. Finally, show the existence of a smaller eigenvalue leads to a contradiction.

Exercise 7.3.12 Determine the best constant in the Poincaré–Friedrichs inequality

$$\|v\|_{L^2(a,b)} \leq c \, \|v'\|_{L^2(a,b)} \quad \forall \, v \in V$$

for each of the following spaces:

$$V = H^1_0(a,b),$$
$$V = H^1_{(0}(a,b) = \{v \in H^1(a,b) \mid v(a) = 0\},$$
$$V = H^1_{0)}(a,b) = \{v \in H^1(a,b) \mid v(b) = 0\}.$$

Exercise 7.3.13 Let $\Omega \subset \Pi_{i=1}^d (0, l_i)$ be a Lipschitz domain in $\mathbb{R}^d$. Recall the Poincaré inequality:

$$\|v\|_{L^2(\Omega)} \leq c \, \|\nabla v\|_{L^2(\Omega)} \quad \forall \, v \in H^1_0(\Omega).$$

Show by an elementaryu argument that we may take $c = (1/\sqrt{2}) \, \min_{1 \leq i \leq d} l_i$. Show with a more delicate argument that $c = \pi^{-1} \left(\sum_{i=1}^d l_i^{-2} \right)^{-1/2}$.

Hint: Extend v by value zero on $\Omega_0 \backslash \Omega$, where $\Omega_0 = \Pi_{i=1}^d (0, l_i)$. Then $v \in H^1_0(\Omega_0)$, and it is sufficient to prove the results for $\Omega = \Omega_0$.

Exercise 7.3.14 In Exercises 7.3.11 and 7.3.12, the best constant of an inequality is related to a linear eigenvalue boundary value problem. In some other applications, we need the best constant of an inequality, which can be found or estimated by solving a linear elliptic boundary value problem. Keeping the notations of Exercise 7.3.11, we have the Sobolev inequality

$$\|v\|_{L^1(\Gamma_1)} \leq c \, \|\nabla v\|_{L^2(\Omega)} \quad \forall \, v \in H^1_{\Gamma_2}(\Omega).$$

The best constant c_0 of the inequality can be characterized by the expression

$$c_0 = \sup\{\|v\|_{L^1(\Gamma_1)} / \|\nabla v\|_{L^2(\Omega)} \mid v \in H^1_{\Gamma_2}(\Omega)\}.$$

Show that

$$c_0 = \|\nabla u\|_{L^2(\Omega)} = \|u\|^{1/2}_{L^1(\Gamma_1)},$$

where, u is the solution of the problem

$$u \in H^1_{\Gamma_2}(\Omega), \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_{\Gamma_1} v \, ds \quad \forall \, v \in H^1_{\Gamma_2}(\Omega).$$

Hint: Use the result of Exercise 7.1.5 (see [110, Section 1.5] or [108]).

7.4 Characterization of Sobolev spaces via the Fourier transform

When $\Omega = \mathbb{R}^d$, it is possible to define Sobolev spaces $H^k(\mathbb{R}^d)$ by using the Fourier transform. All the functions in this section are allowed to be complex-valued. A review on the theory of the Fourier transform is given in Section 4.2. For $v \in L^1(\mathbb{R}^d)$, its Fourier transform is

$$\mathcal{F}v(\boldsymbol{\xi}) = \frac{1}{(2\,\pi)^{d/2}} \int_{\mathbb{R}^d} \exp(-i\,\boldsymbol{x}\cdot\boldsymbol{\xi})\, v(\boldsymbol{x})\, dx,$$

and the inverse Fourier transform is

$$\mathcal{F}^{-1}v(\boldsymbol{x}) = \frac{1}{(2\,\pi)^{d/2}} \int_{\mathbb{R}^d} \exp(i\,\boldsymbol{x}\cdot\boldsymbol{\xi})\, v(\boldsymbol{\xi})\, d\xi.$$

By (4.2.16) and (4.2.17), these two formulas are valid also for $v \in L^2(\mathbb{R}^d)$, in the sense given in (4.2.14) and (4.2.15). We recall the important property

$$\mathcal{F}(\partial^\alpha v)(\boldsymbol{\xi}) = (i\,\boldsymbol{\xi})^\alpha \mathcal{F}v(\boldsymbol{\xi}) \quad \text{if } \partial^\alpha v \in L^2(\mathbb{R}^d).$$

It is then straightforward to show the next result.

Theorem 7.4.1 *A function $v \in L^2(\mathbb{R}^d)$ belongs to $H^k(\mathbb{R}^d)$ if and only if $(1+|\boldsymbol{\xi}|^2)^{k/2}\mathcal{F}v \in L^2(\mathbb{R}^d)$. Moreover, there exist $c_1, c_2 > 0$ such that*

$$c_1\|v\|_{H^k(\mathbb{R}^d)} \le \|(1+|\boldsymbol{\xi}|^2)^{k/2}\mathcal{F}v\|_{L^2(\mathbb{R}^d)} \le c_2\|v\|_{H^k(\mathbb{R}^d)} \quad \forall\, v \in H^k(\mathbb{R}^d). \tag{7.4.1}$$

Thus we see that $\|(1+|\boldsymbol{\xi}|^2)^{k/2}\mathcal{F}v\|_{L^2(\mathbb{R}^d)}$ is a norm on $H^k(\mathbb{R}^d)$, which is equivalent to the canonical norm $\|v\|_{H^k(\mathbb{R}^d)}$. Other equivalent norms can also be used, e.g., $\|(1+|\boldsymbol{\xi}|^k)\mathcal{F}v\|_{L^2(\mathbb{R}^d)}$ or $\|(1+|\boldsymbol{\xi}|)^k\mathcal{F}v\|_{L^2(\mathbb{R}^d)}$. So it is equally well to define the space $H^k(\mathbb{R}^d)$ as

$$H^k(\mathbb{R}^d) = \{v \in L^2(\mathbb{R}^d) \mid (1+|\boldsymbol{\xi}|^2)^{k/2}\mathcal{F}v \in L^2(\mathbb{R}^d)\}.$$

We notice that in this equivalent definition, there is no need to assume k to be an integer. It is natural to define Sobolev spaces of any (positive) order $s \ge 0$:

$$H^s(\mathbb{R}^d) = \{v \in L^2(\mathbb{R}^d) \mid (1+|\boldsymbol{\xi}|^2)^{s/2}\mathcal{F}v \in L^2(\mathbb{R}^d)\} \tag{7.4.2}$$

with the norm

$$\|v\|_{H^s(\mathbb{R}^d)} = \|(1+|\boldsymbol{\xi}|^2)^{s/2}\mathcal{F}v\|_{L^2(\mathbb{R}^d)}$$

and the inner product

$$(u,v)_{H^s(\mathbb{R}^d)} = \int_{\mathbb{R}^d} (1+|\boldsymbol{\xi}|^2)^s \mathcal{F}u(\boldsymbol{\xi})\,\overline{\mathcal{F}v(\boldsymbol{\xi})}\, d\xi.$$

In particular, when $s = 0$, we recover the $L^2(\mathbb{R}^d)$ space: $H^0(\mathbb{R}^d) = L^2(\mathbb{R}^d)$.

Actually, the definition of the Fourier transform can be extended to distributions of slow growth which are continuous linear functionals on smooth functions decaying sufficiently rapidly at infinity (see Section 4.2 or [210] for detailed development of this subject). Then we can define the Sobolev space $H^s(\mathbb{R}^d)$ for negative index s to be the set of the distributions v of slow growth such that

$$\|v\|_{H^s(\mathbb{R}^d)} = \|(1 + |\boldsymbol{\xi}|^2)^{s/2}\mathcal{F}v\|_{L^2(\mathbb{R}^d)} < \infty.$$

We can combine the extension theorem, the approximation theorems and the Fourier transform characterization of Sobolev spaces to prove some properties of Sobolev spaces over bounded Lipschitz domains.

Example 7.4.2 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Assume $k > d/2$. Prove $H^k(\Omega) \hookrightarrow C(\overline{\Omega})$, i.e.,

$$\|v\|_{C(\overline{\Omega})} \le c\,\|v\|_{H^k(\Omega)} \quad \forall\, v \in H^k(\Omega). \tag{7.4.3}$$

Proof. Step 1. We prove (7.4.3) for $\Omega = \mathbb{R}^d$ and $v \in C_0^\infty(\mathbb{R}^d)$. We have

$$v(\boldsymbol{x}) = \frac{1}{(2\,\pi)^{d/2}} \int_{\mathbb{R}^d} \exp(i\,\boldsymbol{x}\cdot\boldsymbol{\xi})\mathcal{F}v(\boldsymbol{\xi})\,d\xi.$$

Thus

$$\begin{aligned}
|v(\boldsymbol{x})| &\le c \int_{\mathbb{R}^d} |\mathcal{F}v(\boldsymbol{\xi})|\,d\xi \\
&= c \int_{\mathbb{R}^d} \frac{(1 + |\boldsymbol{\xi}|^2)^{k/2}|\mathcal{F}v(\boldsymbol{\xi})|}{(1 + |\boldsymbol{\xi}|^2)^{k/2}} d\xi \\
&\le c \left[\int_{\mathbb{R}^d} (1 + |\boldsymbol{\xi}|^2)^{-k} d\xi\right]^{1/2} \left[\int_{\mathbb{R}^d} (1 + |\boldsymbol{\xi}|^2)^k |\mathcal{F}v(\boldsymbol{\xi})|^2 d\xi\right]^{1/2} \\
&= c \left[\int_{\mathbb{R}^d} (1 + |\boldsymbol{\xi}|^2)^k |\mathcal{F}v(\boldsymbol{\xi})|^2 d\xi\right]^{1/2},
\end{aligned}$$

where we used the fact that

$$\int_{\mathbb{R}^d} \frac{1}{(1 + |\boldsymbol{\xi}|^2)^k}\,d\xi = \omega_{d-1} \int_0^\infty \frac{r^{d-1}}{(1 + r^2)^k}\,dr < \infty \text{ if and only if } k > d/2$$

(ω_{d-1} denotes the $(d-1)$-dimensional Lebesgue measure of the unit sphere in $\mathbb{R}^d$). Hence,

$$\|v\|_{C(\mathbb{R}^d)} \le c\,\|v\|_{H^k(\mathbb{R}^d)} \quad \forall\, v \in C_0^\infty(\mathbb{R}^d).$$

Step 2. Since $C_0^\infty(\mathbb{R}^d)$ is dense in $H^k(\mathbb{R}^d)$, the relation (7.4.3) holds for any $v \in H^k(\mathbb{R}^d)$.

STEP 3. We now use the extension theorem. For any $v \in H^k(\Omega)$, we can extend it to $Ev \in H^k(\mathbb{R}^d)$ with

$$\|Ev\|_{H^k(\mathbb{R}^d)} \le c\,\|v\|_{H^k(\Omega)}.$$

Therefore,

$$\|v\|_{C(\overline{\Omega})} \le \|Ev\|_{C(\mathbb{R}^d)} \le c\,\|Ev\|_{H^k(\mathbb{R}^d)} \le c\,\|v\|_{H^k(\Omega)}.$$

Thus we have proved (7.4.3). □

Example 7.4.3 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Then

$$\|v\|_{C(\overline{\Omega})} \le c\,\|v\|^{1/2}_{H^d(\Omega)}\|v\|^{1/2}_{L^2(\Omega)} \quad \forall\, v \in H^d(\Omega). \tag{7.4.4}$$

Proof. As in the previous example, it is sufficient to show (7.4.4) for the case $\Omega = \mathbb{R}^d$ and $v \in C_0^\infty(\Omega)$. For any $\lambda > 0$,

$$\begin{aligned}
|v(\boldsymbol{x})|^2 &\le c\left[\int_{\mathbb{R}^d} |\mathcal{F}v(\boldsymbol{\xi})|\,d\xi\right]^2 \\
&= c\left[\int_{\mathbb{R}^d} \frac{(1+\lambda\,|\boldsymbol{\xi}|^2)^{d/2}|\mathcal{F}v(\boldsymbol{\xi})|}{(1+\lambda\,|\boldsymbol{\xi}|^2)^{d/2}}\,d\xi\right]^2 \\
&\le c\int_{\mathbb{R}^d}(1+\lambda\,|\boldsymbol{\xi}|^2)^d|\mathcal{F}v(\boldsymbol{\xi})|^2 d\xi \int_{\mathbb{R}^d}(1+\lambda\,|\boldsymbol{\xi}|^2)^{-d} d\xi \\
&\le c\,\frac{1}{\lambda^{d/2}}\int_{\mathbb{R}^d}\left[|\mathcal{F}v(\boldsymbol{\xi})|^2 + \lambda^d(1+|\boldsymbol{\xi}|^2)^d|\mathcal{F}v(\boldsymbol{\xi})|^2\right] d\xi \\
&= c\left[\lambda^{-d/2}\|v\|^2_{L^2(\mathbb{R}^d)} + \lambda^{d/2}\|v\|^2_{H^d(\mathbb{R}^d)}\right].
\end{aligned}$$

Taking

$$\lambda = \left[\|v\|_{L^2(\mathbb{R}^d)}/\|v\|_{H^d(\mathbb{R}^d)}\right]^{2/d},$$

we then get the required inequality. □

Exercise 7.4.1 Let $v(x)$ be a step function defined by: $v(x) = 1$ for $x \in [0,1]$, and $v(x) = 0$ for $x \notin [0,1]$. Find the range of s for which $v \in H^s(\mathbb{R})$.

Exercise 7.4.2 As a part of a proof of Theorem 7.4.1, show that there exist constants $c_1, c_2 > 0$ such that

$$c_1\left(\sum_{|\alpha|\le k}|\boldsymbol{\xi}^\alpha|^2\right)^{1/2} \le (1+|\boldsymbol{\xi}|^2)^{k/2} \le c_2\left(\sum_{|\alpha|\le k}|\boldsymbol{\xi}^\alpha|^2\right)^{1/2} \quad \forall\,\boldsymbol{\xi} \in \mathbb{R}^d.$$

Exercise 7.4.3 Provide a detailed argument for Step 2 in the proof of Example 7.4.2.

Exercise 7.4.4 From the Sobolev embedding theorem, we know that $H^s(\Omega) \hookrightarrow C(\overline{\Omega})$ whenever $s > d/2$. For $s \in (d/2, d)$, derive an inequality of the form (7.4.4) for $v \in H^s(\Omega)$.

Exercise 7.4.5 Let $\sigma \in [0, 1]$. Show the Sobolev space interpolation inequality:

$$\|v\|_{H^\sigma(\mathbb{R}^d)} \leq \|v\|^{\sigma}_{H^1(\mathbb{R}^d)} \|v\|^{1-\sigma}_{L^2(\mathbb{R}^d)} \quad \forall\, v \in H^1(\mathbb{R}^d).$$

More generally, let $s_0 < s_1$ be two real numbers, and let $s_\sigma = \sigma\, s_1 + (1 - \sigma)\, s_0$ for some $\sigma \in [0, 1]$. Then the following interpolation inequality holds:

$$\|v\|_{H^{s_\sigma}(\mathbb{R}^d)} \leq \|v\|^{\sigma}_{H^{s_1}(\mathbb{R}^d)} \|v\|^{1-\sigma}_{H^{s_0}(\mathbb{R}^d)} \quad \forall\, v \in H^{s_1}(\mathbb{R}^d).$$

Exercise 7.4.6 Show that the formula

$$\|v\|^*_{H^s(\mathbb{R}^d)} = \|(1 + |\boldsymbol{\xi}|^s)\mathcal{F}v\|_{L^2(\mathbb{R}^d)}$$

defines a norm on $H^s(\mathbb{R}^d)$ and this norm is equivalent to the norm $\|v\|_{H^s(\mathbb{R}^d)}$.

7.5 Periodic Sobolev spaces

When working with an equation defined over the boundary of a bounded and simply-connected region in the plane, the functions being discussed are periodic. Therefore, it is useful to consider Sobolev spaces of such functions. Since periodic functions are often discussed with reference to their Fourier series expansion, we use this expansion to discuss Sobolev spaces of such functions.

From Example 1.3.15, we write the Fourier series of $\varphi \in L^2(0, 2\pi)$ as

$$\varphi(x) = \sum_{m=-\infty}^{\infty} a_m \psi_m(x)$$

with

$$\psi_m(x) = \frac{1}{\sqrt{2\pi}}\, e^{imx}, \quad m = 0, \pm 1, \pm 2, \ldots$$

forming an orthonormal basis of $L^2(0, 2\pi)$. The Fourier coefficients are given by

$$a_m = (\varphi, \psi_m) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} \varphi(x)\, e^{-imx} dx.$$

Convergence of the Fourier series was discussed earlier in Section 4.1. Also, for a non-negative integer k, recall that $C^k_p(2\pi)$ denotes the space of all periodic functions on $(-\infty, \infty)$ with period 2π that are also k-times continuously differentiable.

Definition 7.5.1 *For an integer $k \geq 0$, $H^k(2\pi)$ is defined to be the closure of $C_p^k(2\pi)$ under the inner product norm*

$$\|\varphi\|_{H^k} \equiv \left[\sum_{j=0}^{k} \left\| \varphi^{(j)} \right\|_{L^2}^2 \right]^{1/2}.$$

For arbitrary real $s \geq 0$, $H^s(2\pi)$ can be obtained as in earlier sections, with the formulas of Section 7.4 being closest to the discussion given below, especially (7.4.2).

The following can be shown without too much difficulty; e.g. see [149, Chap. 8].

Theorem 7.5.2 *For $s \in \mathbb{R}$, $H^s(2\pi)$ is the set of all series*

$$\varphi(x) = \sum_{m=-\infty}^{\infty} a_m \psi_m(x) \tag{7.5.1}$$

for which

$$\|\varphi\|_{*,s}^2 \equiv |a_0|^2 + \sum_{|m|>0} |m|^{2s} |a_m|^2 < \infty. \tag{7.5.2}$$

Moreover, the norm $\|\varphi\|_{*,s}$ is equivalent to the standard Sobolev norm $\|\varphi\|_{H^s}$ for $\varphi \in H^s(2\pi)$.

The norm $\|\cdot\|_{*,s}$ is induced by the inner product

$$(\varphi, \rho)_{*,s} \equiv a_0 \overline{b}_0 + \sum_{|m|>0} |m|^{2s} a_m \overline{b}_m$$

where $\varphi = \sum a_m \psi_m$ and $\rho = \sum b_m \psi_m$.

For $s < 0$, the space $H^s(2\pi)$ contains series that are divergent according to most usual definitions of convergence. These new "functions" (7.5.1) are referred to as both *generalized functions* and *distributions*. One way of giving meaning to these new functions is to introduce the concept of *distributional derivative*, which generalizes the derivative in ordinary sense and the weak derivative introduced in Section 7.1. With the ordinary differentiation operator $\mathcal{D} \equiv d/dt$, we have

$$\mathcal{D}\varphi(t) \equiv \frac{d\varphi(t)}{dt} = i \sum_{m=-\infty}^{\infty} m a_m \psi_m(t) \tag{7.5.3}$$

and $\mathcal{D} : H^s(2\pi) \to H^{s-1}(2\pi)$, $s \geq 1$. The distributional derivative gives meaning to differentiation of periodic functions in $L^2(0, 2\pi)$, and also to repeated differentiation of generalized functions. To prove that there exists a unique such extension of the definition of $\mathcal{D}$, proceed as follows.

Introduce the space $\mathbb{T}$ of all trigonometric polynomials:

$$\mathbb{T} = \left\{ \varphi \equiv \sum_{m=-n}^{n} a_m \psi_m \;\middle|\; a_m \in \mathbb{C},\ |m| \le n,\ n = 0, 1, \dots \right\}. \tag{7.5.4}$$

It is straightforward to show this is a dense subspace of $H^s(2\pi)$ for arbitrary s, meaning that when using the norm (7.5.2), the closure of $\mathbb{T}$ equals $H^s(2\pi)$. Considering $\mathbb{T}$ as a subspace of $H^s(2\pi)$, define $\mathcal{D} : \mathbb{T} \to H^{s-1}(2\pi)$ by

$$\mathcal{D}\varphi = \varphi', \quad \varphi \in \mathbb{T}. \tag{7.5.5}$$

This is a bounded operator; and using the representation of φ in (7.5.4), it is straightforward to show

$$\|\mathcal{D}\| = 1.$$

Since $\mathbb{T}$ is dense in $H^s(2\pi)$, and since $\mathcal{D} : \mathbb{T} \subset H^s(2\pi) \to H^{s-1}(2\pi)$ is bounded, we have that there is a unique bounded extension of $\mathcal{D}$ to all of $H^s(2\pi)$; see Theorem 2.4.1. We will retain the notation $\mathcal{D}$ for the extension. Combining the representation of $\varphi \in \mathbb{T}$ in (7.5.4) with the definition (7.5.5), and using the continuity of the extension $\mathcal{D}$, the formula (7.5.5) remains valid for any $\varphi \in H^s(2\pi)$ for all s.

Example 7.5.3 Define

$$\varphi(t) = \begin{cases} 0, & (2k-1)\pi < t < 2k\pi, \\ 1, & 2k\pi < t < (2k+1)\pi, \end{cases}$$

for all integers k, a so-called "square wave". The Fourier series of this function is given by

$$\varphi(t) = \frac{1}{2} - \frac{i}{\pi} \sum_{k=0}^{\infty} \frac{1}{2k+1} \left[e^{(2k+1)it} - e^{-(2k+1)it} \right], \quad -\infty < t < \infty$$

which converges almost everywhere. Regarding this series as a function defined on $\mathbb{R}$, the distributional derivative of $\varphi(t)$ is

$$\varphi'(t) = \frac{1}{\pi} \sum_{k=0}^{\infty} \left[e^{(2k+1)it} + e^{-(2k+1)it} \right] = \sum_{j=-\infty}^{\infty} (-1)^j \delta(t - \pi j).$$

The function $\delta(t)$ is the *Dirac delta function*, and it is a well-studied linear functional on elements of $H^s(2\pi)$ for $s > 1/2$:

$$\delta[\varphi] \equiv \langle \varphi, \delta \rangle = \varphi(0), \quad \varphi \in H^s(2\pi),\ s > 1/2$$

with $\delta[\varphi]$ denoting the action of δ on φ. □

7.5.1 The dual space

The last example suggests another interpretation of $H^s(2\pi)$ for negative s, that as a dual space. Let ℓ be a bounded linear functional on $H^t(2\pi)$ for some $t \geq 0$, bounded with respect to the norm $\|\cdot\|_t$. Then using the *Riesz representation theorem* (Theorem 2.5.8), we have a unique element $\eta_\ell \in H^{-t}(2\pi)$, $\eta_\ell = \sum_{m=-\infty}^{\infty} b_m \psi_m$, with

$$\ell[\varphi] \equiv \langle \varphi, \eta_\ell \rangle = \sum_{m=-\infty}^{\infty} a_m \overline{b}_m \quad \text{for } \varphi = \sum_{m=-\infty}^{\infty} a_m \psi_m \in H^t(2\pi). \tag{7.5.6}$$

It is also straightforward to show that when given two such linear functionals, say ℓ_1 and ℓ_2, we have

$$\eta_{c_1\ell_1 + c_2\ell_2} = \overline{c}_1 \eta_{\ell_1} + \overline{c}_2 \eta_{\ell_2}$$

for all scalars c_1, c_2. Moreover,

$$|\langle \varphi, \eta_\ell \rangle| \leq \|\varphi\|_{*,t} \|\eta_\ell\|_{*,-t}$$

and

$$\|\ell\| = \|\eta_\ell\|_{*,-t}.$$

The space $H^{-t}(2\pi)$ can be used to represent the space of bounded linear functionals on $H^t(2\pi)$, and it is usually called the *dual space* for $H^t(2\pi)$. In this framework, we are regarding $H^0(2\pi) \equiv L^2(0, 2\pi)$ as self-dual. The evaluation of linear functionals on $H^t(2\pi)$, as in (7.5.6), can be considered as a bilinear function defined on $H^t(2\pi) \times H^{-t}(2\pi)$; and in that case,

$$|\langle \varphi, \eta \rangle| \leq \|\varphi\|_{*,t} \|\eta_\ell\|_{*,-t}, \quad \varphi \in H^t(2\pi),\ \eta \in H^{-t}(2\pi). \tag{7.5.7}$$

Define $b : H^t(2\pi) \times L^2(0, 2\pi) \to \mathbb{C}$ by

$$b(\varphi, \psi) = (\varphi, \psi), \quad \varphi \in H^t(2\pi),\ \psi \in L^2(0, 2\pi) \tag{7.5.8}$$

using the usual inner product of $L^2(0, 2\pi)$. Then the bilinear *duality pairing* $\langle \cdot, \cdot \rangle$ is the unique bounded extension of $b(\cdot, \cdot)$ to $H^t(2\pi) \times H^{-t}(2\pi)$, when $L^2(0, 2\pi)$ is regarded as a dense subspace of $H^{-t}(2\pi)$. For a more extensive discussion of this topic with much greater generality, see Aubin [27, Chapter 3].

We have considered $\langle \cdot, \cdot \rangle$ as defined on $H^t(2\pi) \times H^{-t}(2\pi)$ with $t \geq 0$. But we can easily extend this definition to allow $t < 0$. Using (7.5.6), we define

$$\langle \varphi, \eta \rangle = \sum_{m=-\infty}^{\infty} a_m \overline{b}_m \tag{7.5.9}$$

for $\varphi = \sum_{m=-\infty}^{\infty} a_m \psi_m$ in $H^t(2\pi)$ and $\eta = \sum_{m=-\infty}^{\infty} b_m \psi_m$ in $H^{-t}(2\pi)$, for any real number t. The bound (7.5.7) is also still valid. This extension of $\langle \cdot, \cdot \rangle$ is merely a statement that the dual space for $H^t(2\pi)$ with $t < 0$ is just $H^{-t}(2\pi)$.

7.5.2 Embedding results

We give another variant of the Sobolev embedding theorem.

Proposition 7.5.4 *Let $s > k + 1/2$ for some integer $k \geq 0$, and let $\varphi \in H^s(2\pi)$. Then $\varphi \in C_p^k(2\pi)$.*

Proof. We give a proof for only the case $k = 0$, as the general case is quite similar. We show the Fourier series (7.5.1) for φ is absolutely and uniformly convergent on $[0, 2\pi]$; and it then follows by standard arguments that φ is continuous and periodic. From the definition (7.5.1), and by using the Cauchy-Schwarz inequality,

$$\begin{aligned} |\varphi(s)| &\leq \sum_{m=-\infty}^{\infty} |a_m| \\ &= |a_0| + \sum_{|m|>0} |m|^{-s} |m|^s |a_m| \\ &\leq |a_0| + \sqrt{\sum_{|m|>0} |m|^{-2s}} \sqrt{\sum_{|m|>0} |m|^{2s} |a_m|^2}. \end{aligned}$$

Denoting $\zeta(r)$ the *zeta* function:

$$\zeta(r) = \sum_{m=1}^{\infty} \frac{1}{m^r}, \quad r > 1,$$

we then have

$$|\varphi(s)| \leq |a_0| + \sqrt{2\zeta(2s)}\, \|\varphi\|_s. \tag{7.5.10}$$

By standard arguments on the convergence of infinite series, (7.5.10) implies that the Fourier series (7.5.1) for φ is absolutely and uniformly convergent. In addition,

$$\|\varphi\|_\infty \leq \left[1 + \sqrt{2\zeta(2s)}\right] \|\varphi\|_{*,s}, \quad \varphi \in H^s(2\pi). \tag{7.5.11}$$

Thus the identity mapping from $H^s(2\pi)$ into $C_p(2\pi)$ is bounded. □

The proof of the following result is left as Exercise 7.5.1.

Proposition 7.5.5 *Let $s > t$. Then $H^s(2\pi)$ is dense in $H^t(2\pi)$, and the identity mapping*

$$I : H^s(2\pi) \to H^t(2\pi), \quad I(\varphi) \equiv \varphi \quad \text{for } \varphi \in H^s(2\pi)$$

is a compact operator.

7.5.3 Approximation results

When integrals of periodic functions are approximated numerically, the trapezoidal rule is the method of choice in most cases. Let us explain why. Suppose the integral to be evaluated is

$$I(\varphi) = \int_0^{2\pi} \varphi(x)\,dx$$

with $\varphi \in H^s(2\pi)$ and $s > 1/2$. The latter assumption guarantees φ is continuous so that evaluation of $\varphi(x)$ makes sense for all x. For an integer $k \geq 1$, let $h = 2\pi/k$ and write the trapezoidal rule as

$$T_k(\varphi) = h \sum_{j=1}^{k} \varphi(jh). \tag{7.5.12}$$

We give a nonstandard error bound, but one that shows the rapid rate of convergence for smooth periodic functions.

Proposition 7.5.6 *Assume $s > 1/2$, and let $\varphi \in H^s(2\pi)$. Then*

$$|I(\varphi) - T_k(\varphi)| \leq \frac{\sqrt{4\pi\zeta(2s)}}{k^s}\|\varphi\|_s, \quad k \geq 1. \tag{7.5.13}$$

Proof. We begin with the following result, on the application of the trapezoidal rule to e^{imx}.

$$T_k(e^{imx}) = \begin{cases} 2\pi, & m = jk,\ j = 0, \pm 1, \pm 2, \ldots, \\ 0, & \text{otherwise.} \end{cases} \tag{7.5.14}$$

Using it, and applying T_k to the Fourier series representation (7.5.1) of $\varphi(s)$, we have

$$I(\varphi) - T_k(\varphi) = -\sqrt{2\pi} \sum_{|m|>0} a_{km} = -\sqrt{2\pi} \sum_{|m|>0} a_{km}(km)^s (km)^{-s}.$$

Applying the Cauchy-Schwarz inequality to the last sum,

$$\begin{aligned} |I(\varphi) - T_k(\varphi)| &\leq \sqrt{2\pi} \left[\sum_{|m|>0} |a_{km}|^2 (km)^{2s}\right]^{1/2} \left[\sum_{|m|>0} (km)^{-2s}\right]^{1/2} \\ &\leq \sqrt{2\pi}\,\|\varphi\|_s\, k^{-s}\sqrt{2\zeta(2s)}. \end{aligned}$$

This completes the proof. □

Recall the discussion of the trigonometric interpolation polynomial $\mathcal{I}_n\varphi$ in Chapter 3, and in particular the theoretical error bound of Theorem 3.7.1. We give another error bound for this interpolation. A complete derivation of it can be found in [151].

Theorem 7.5.7 *Let $s > 1/2$, and let $\varphi \in H^s(2\pi)$. Then for $0 \le r \le s$,*

$$\|\varphi - \mathcal{I}_n\varphi\|_r \le \frac{c}{n^{s-r}}\|\varphi\|_s, \quad n \ge 1. \tag{7.5.15}$$

The constant c depends on s and r only.

Proof. The proof is based on using the Fourier series representation (7.5.1) of φ to obtain a Fourier series for $\mathcal{I}_n\varphi$. This is then subtracted from that for φ, and the remaining terms are bounded to give the result (7.5.15). For details, see [151]. This is only a marginally better result than Theorem 3.7.1, but it is an important tool when doing error analyses of numerical methods in the Sobolev spaces $H^s(2\pi)$. □

7.5.4 An illustrative example of an operator

To illustrate the usefulness of the Sobolev spaces $H^s(2\pi)$, we consider the following important integral operator:

$$\mathcal{A}\varphi(x) = -\frac{1}{\pi}\int_0^{2\pi} \varphi(y)\log\left|2e^{-1/2}\sin\left(\frac{x-y}{2}\right)\right| dy, \quad -\infty < x < \infty, \tag{7.5.16}$$

for $\varphi \in L^2(0, 2\pi)$. It plays a critical role in the study of boundary integral equation reformulations of Laplace's equation in the plane. Using results from the theory of functions of a complex variable, one can show that

$$\mathcal{A}\varphi(t) = a_0\psi_0(t) + \sum_{|m|>0} \frac{a_m}{|m|}\psi_m(t) \tag{7.5.17}$$

where $\varphi = \sum_{m=-\infty}^{\infty} a_m\psi_m$. In turn, this implies that

$$\mathcal{A} : H^s(2\pi) \underset{onto}{\overset{1-1}{\to}} H^{s+1}(2\pi), \quad s \ge 0 \tag{7.5.18}$$

and

$$\|\mathcal{A}\| = 1.$$

The definition of $\mathcal{A}$ as an integral operator in (7.5.16) requires that the function φ be a function to which integration can be applied. However, the formula (7.5.17) permits us to extend the domain for $\mathcal{A}$ to any Sobolev space $H^s(2\pi)$ with $s < 0$. This is important in that one important approach to the numerical analysis of the equation $\mathcal{A}\varphi = f$ requires $\mathcal{A}$ to be regarded as an operator from $H^{-1/2}(2\pi)$ to $H^{1/2}(2\pi)$.

We will return to the study and application of $\mathcal{A}$ in Chapter 13; and it is discussed at length in [18, Chap. 7] and [206].

7.5.5 Spherical polynomials and spherical harmonics

The Fourier series can be regarded as an expansion of functions defined on the unit circle in the plane. For functions defined on the unit sphere U in $\mathbb{R}^3$, the analogue of the Fourier series is the *Laplace expansion*, and it uses *spherical harmonics* as the generalizations of the trigonometric functions. We begin by first considering *spherical polynomials*, and then we introduce the spherical harmonics and Laplace expansion. Following the tradition in the literature on the topic, we use (x, y, z) for a generic point in $\mathbb{R}^3$ in this subsection.

Definition 7.5.8 *Consider an arbitrary polynomial in (x, y, z) of degree N, say*

$$p(x, y, z) = \sum_{\substack{i,j,k\geq 0 \\ i+j+k\leq N}} a_{i,j,k} x^i y^j z^k, \tag{7.5.19}$$

and restrict (x, y, z) to lie on the unit sphere U. The resulting function is called a spherical polynomial *of degree $\leq N$. Denote by $\mathbb{S}_N$ the set of all such spherical polynomials of degree $\leq N$.*

Spherical polynomials are the analogues of the trigonometric polynomials, which can be obtained by replacing (x, y) in

$$\sum_{\substack{i,j\geq 0 \\ i+j\leq N}} a_{i,j} x^i y^j$$

with $(\cos\theta, \sin\theta)$. Note that a polynomial $p(x, y, z)$ may reduce to an expression of lower degree. For example, $p(x, y, z) = x^2 + y^2 + z^2$ reduces to 1 when $(x, y, z) \in U$.

An alternative way of obtaining polynomials on U is to begin with *homogeneous harmonic polynomials.*

Definition 7.5.9 *Let $p = p(x, y, z)$ be a polynomial of degree n which satisfies Laplace's equation,*

$$\Delta p(x, y, z) \equiv \frac{\partial^2 p}{\partial x^2} + \frac{\partial^2 p}{\partial y^2} + \frac{\partial^2 p}{\partial z^2} = 0, \quad (x, y, z) \in \mathbb{R}^3,$$

and further, let p be homogeneous of degree n:

$$p(tx, ty, tz) = t^n p(x, y, z), \quad -\infty < t < \infty, \ (x, y, z) \in \mathbb{R}^3.$$

Restrict all such polynomials to U. Such functions are called spherical harmonics *of degree n.*

As examples of spherical harmonics, we have the following.

1. $n = 0$: $p(x, y, z) = 1$,
2. $n = 1$: $p(x, y, z) = x,\ y,\ z$,
3. $n = 2$: $p(x, y, z) = xy,\ xz,\ yz,\ x^2 + y^2 - 2z^2,\ x^2 + z^2 - 2y^2$,

where in all cases, we use $(x, y, z) = (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta)$. Nontrivial linear combinations of spherical harmonics of a given degree are again spherical harmonics of that same degree. For example,

$$p(x, y, z) = x + y + z$$

is also a spherical harmonic of degree 1. The number of linearly independent spherical harmonics of degree n is $2n + 1$; and thus the above sets are maximal independent sets for each of the given degrees $n = 0, 1, 2$.

Define $\widehat{\mathbb{S}}_N$ to be the smallest vector space to contain all of the spherical harmonics of degree $n \le N$. Alternatively, $\widehat{\mathbb{S}}_N$ is the set of all finite linear combinations of spherical harmonics of all possible degrees $n \le N$. Then it can be shown that

$$\mathbb{S}_N = \widehat{\mathbb{S}}_N \tag{7.5.20}$$

and

$$\dim \mathbb{S}_N = (N + 1)^2. \tag{7.5.21}$$

Below, we introduce a basis for $\mathbb{S}_N$. See MacRobert [161, Chap. 7] for a proof of these results.

There are well-known formulas for spherical harmonics, and we will make use of some of them in working with spherical polynomials. The subject of spherical harmonics is quite large, and we can only touch on a few small parts of it. For further study, see the classical book [161] by T. MacRobert. The spherical harmonics of degree n are the analogues of the trigonometric functions $\cos n\theta$ and $\sin n\theta$, which are restrictions to the unit circle of the homogeneous harmonic polynomials

$$r^n \cos(n\theta), \quad r^n \sin(n\theta)$$

written in polar coordinates form.

The standard basis for spherical harmonics of degree n is

$$\begin{aligned} S_n^1(x, y, z) &= c_n L_n(\cos\theta), \\ S_n^{2m}(x, y, z) &= c_{n,m} L_n^m(\cos\theta)\cos(m\phi), \\ S_n^{2m+1}(x, y, z) &= c_{n,m} L_n^m(\cos\theta)\sin(m\phi), \quad m = 1, \ldots, n, \end{aligned} \tag{7.5.22}$$

with $(x, y, z) = (\cos\phi \sin\theta, \sin\phi \sin\theta, \cos\theta)$. In this formula, $L_n(t)$ is a *Legendre polynomial* of degree n,

$$L_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} \left[(t^2 - 1)^n\right] \tag{7.5.23}$$

and $L_n^m(t)$ is an *associated Legendre function*,

$$L_n^m(t) = (-1)^m (1-t^2)^{m/2} \frac{d^m}{dt^m} L_n(t), \quad 1 \le m \le n. \tag{7.5.24}$$

The constants in (7.5.22) are given by

$$c_n = \sqrt{\frac{2n+1}{4\pi}}, \quad c_{n,m} = \sqrt{\frac{2n+1}{2\pi} \frac{(n-m)!}{(n+m)!}}.$$

We will occasionally denote the Legendre polynomial L_n by L_n^0, to simplify referring to these Legendre functions.

The standard inner product on $L^2(U)$ is given by

$$(f,g) = \int_U f(Q)\, g(Q)\, dS_Q.$$

Using this definition, we can verify that the functions of (7.5.22) satisfy

$$(S_n^k, S_q^p) = \delta_{n,q} \delta_{k,p}$$

for $n, q = 0, 1, \ldots$ and $1 \le k \le 2n+1$, $1 \le p \le 2q+1$. The set of functions

$$\{S_n^k \mid 1 \le k \le 2n+1,\ 0 \le n \le N\}$$

is an orthonormal basis for $\mathbb{S}_N$. To avoid some double summations, we will sometimes write this basis for $\mathbb{S}_N$ as

$$\{\Psi_1, \ldots, \Psi_{d_N}\} \tag{7.5.25}$$

with $d_N = (N+1)^2$ the dimension of the subspace.

The set $\{S_n^k \mid 1 \le k \le 2n+1,\ 0 \le n < \infty\}$ of spherical harmonics is an orthonormal basis for $L^2(U)$, and it leads to the expansion formula

$$g(Q) = \sum_{n=0}^{\infty} \sum_{k=1}^{2n+1} (g, S_n^k)\, S_n^k(Q), \quad g \in L^2(U). \tag{7.5.26}$$

This is called the *Laplace expansion* of the function g, and it is the generalization to $L^2(U)$ of the Fourier series on the unit circle in the plane. The function $g \in L^2(U)$ if and only if

$$\|g\|_{L^2}^2 = \sum_{n=0}^{\infty} \sum_{k=1}^{2n+1} \left|(g, S_n^k)\right|^2 < \infty. \tag{7.5.27}$$

In analogy with the use of the Fourier series to define the Sobolev spaces $H^s(2\pi)$, we can characterize the Sobolev spaces $H^s(U)$ by using the Laplace expansion.

Of particular interest is the truncation of the series (7.5.26) to terms of degree at most N, to obtain

$$P_N g(Q) = \sum_{n=0}^{N} \sum_{k=1}^{2n+1} (g, S_n^k)\, S_n^k(Q). \tag{7.5.28}$$

This defines the orthogonal projection of $L^2(U)$ onto $\mathbb{S}_N$; and of course, $P_N g \to g$ as $N \to \infty$. Since it is an orthogonal projection on $L^2(U)$, we have $\|P_N\| = 1$ as an operator from $L^2(U)$ in $L^2(U)$. However, we can also regard $\mathbb{S}_N$ as a subset of $C(S)$; and then regarding P_N as a projection from $C(S)$ to $\mathbb{S}_N$, we have

$$\|P_N\| = \left(\sqrt{\frac{8}{\pi}} + \delta_N\right)\sqrt{N} \tag{7.5.29}$$

with $\delta_N \to 0$ as $N \to \infty$. A proof of this formula is quite involved, and we refer the reader to Gronwall [100] and Ragozin [191]. In a later chapter, we use the projection P_N to define a Galerkin method for solving integral equations defined on U, with $\mathbb{S}_N$ as the approximating subspace.

Best approximations

Given $g \in C(U)$, define

$$\rho_N(g) = \inf_{p \in \mathbb{S}_N} \|g - p\|_\infty . \tag{7.5.30}$$

This is called the *minimax error* for the approximation of g by spherical polynomials of degree $\leq N$. Using the Stone-Weierstraß theorem, Theorem 3.1.2, it can be shown that $\rho_N(g) \to 0$ as $N \to \infty$. In the error analysis of numerical methods that use spherical polynomials, it is important to have bounds on the rate at which $\rho_N(g)$ converges to zero. An initial partial result was given by Gronwall [100]; and a much more complete theory was given many years later by Ragozin [190], a special case of which we give here. We first introduce some notation.

For a given positive integer k, let $\partial^k g$ denote an arbitrary k^{th} order derivative of g on U, formed with respect to local surface coordinates on U. (One should consider a set of local patch coordinate systems over U, as in Definition 7.2.13, and Sobolev spaces based on these patches. But what is intended is clear and the present notation is simpler.) Let γ be a real number, $0 < \gamma \leq 1$. Define $C^{k,\gamma}(U)$ to be the set of all functions $g \in C(U)$ for which all of its derivatives $\partial^k g \in C(U)$, with each of these derivatives satisfying a Hölder condition with exponent γ:

$$\left|\partial^k g(P) - \partial^k g(Q)\right| \leq H_{k,\gamma}(g)\, |P - Q|^\gamma, \quad P, Q \in U.$$

Here $|P - Q|$ denotes the usual distance between the two points P and Q. The Hölder constant $H_{k,\gamma}(g)$ is to be uniform over all k^{th} order derivatives of g.

Theorem 7.5.10 *Let $g \in C^{k,\gamma}(U)$. Then there is a sequence of spherical polynomials $\{p_N\}$ for which $\|g - p_N\|_\infty = \rho_N(g)$ and*

$$\rho_N(g) \leq \frac{c_k H_{k,\gamma}(g)}{N^{k+\gamma}}, \quad N \geq 1. \tag{7.5.31}$$

The constant c_k is dependent on only k.

For the case $k = 0$, see Gronwall [100] for a proof; and for the general case, a proof can be found in Ragozin [190, Theorem 3.3]. A different approach to the problem is given by Rustamov [201], and it leads to more general results, involving norms in addition to the uniform norm.

This theorem leads immediately to results on the rate of convergence of the Laplace series expansion of a function $g \in C^{k,\gamma}(U)$, given in (7.5.26). Using the norm of $L^2(U)$, and using the definition of the orthogonal projection $P_N g$ being the best approximation in the inner product norm, we have, for $N \geq 1$,

$$\|g - P_N g\| \leq \|g - p_N\| \leq 4\pi \|g - p_N\|_\infty \leq \frac{4\pi c_{k,\gamma}(g)}{N^{k+\gamma}}. \tag{7.5.32}$$

We can also consider the uniform convergence of the Laplace series. Write

$$\begin{aligned} \|g - P_N g\|_\infty &= \|g - p_N - P_N(g - p_N)\|_\infty \\ &\leq (1 + \|P_N\|) \|g - p_N\|_\infty \\ &\leq c\, N^{-(k+\gamma-1/2)}, \end{aligned} \tag{7.5.33}$$

with the last step using (7.5.29). In particular, if $g \in C^{0,\gamma}(U)$, $\gamma \in (1/2, 1]$, we have uniform convergence of $P_N g$ to g on U. From (7.5.31), the constant c is a multiple of $H_{k,\gamma}(g)$.

No way is known to interpolate with spherical polynomials in a manner that generalizes trigonometric interpolation. For a more complete discussion of this and other problems in working with spherical polynomial approximations, see [18, Section 5.5].

Sobolev spaces on the unit sphere

The function spaces $L^2(U)$ and $C(U)$ are the most widely used function spaces over U. Nevertheless, we need to also introduce the Sobolev spaces $H^r(U)$. There are several equivalent ways to define $H^r(U)$. The standard way is to proceed as in Definition 7.2.13, using local coordinate systems based on a local set of patches covering U and then using Sobolev spaces based on these patches.

Another approach, used less often but possibly more intuitive, is based on the Laplace expansion of a function g defined on the unit sphere U. Recalling the Laplace expansion (7.5.26), define the Sobolev space $H^r(U)$ to be the set of functions whose Laplace expansion satisfies

$$\|g\|_{*,r} \equiv \left[\sum_{n=0}^{\infty} (2n+1)^{2r} \sum_{k=1}^{2n+1} \left|(g, S_n^k)\right|^2\right]^{1/2} < \infty. \tag{7.5.34}$$

This definition can be used for any real number $r \geq 0$. For r a positive integer, the norm $\|g\|_{*,r}$ can be shown to be equivalent to a standard Sobolev norm based on a set of local patch coordinate systems for U.

Exercise 7.5.1 Prove Proposition 7.5.5.

Exercise 7.5.2 Prove that the space $\mathbb{T}$ defined in (7.5.4) is dense in $H^s(2\pi)$ for $-\infty < s < \infty$.

Exercise 7.5.3 Let φ be a continuous periodic function with period 2π. Write Simpson's rule as

$$I(\varphi) = \int_0^{2\pi} \varphi(x)\,dx \approx \frac{h}{3}\left[4\sum_{j=1}^{k}\varphi(x_{2j-1}) + 2\sum_{j=1}^{k}\varphi(x_{2j})\right] \equiv S_{2k}(\varphi),$$

where $h = 2\pi/(2k) = \pi/k$. Show that if $\varphi \in H^s(2\pi)$, $s > 1/2$, then

$$|I(\varphi) - S_{2k}(\varphi)| \leq c\,k^{-s}\|\varphi\|_s.$$

Exercise 7.5.4 For $t > 0$, demonstrate that elements of $H^{-t}(2\pi)$ are indeed bounded linear functionals on $H^t(2\pi)$ when using (7.5.6) to define the linear functional.

7.6 Integration by parts formulas

We comment on the validity of integration by parts formulas. Assume Ω is a Lipschitz domain in $\mathbb{R}^d$ with boundary Γ. Denote $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^T$ the unit outward normal vector on Γ, which is defined almost everywhere giving that Γ is Lipschitz continuous. It is a well-known classical result that

$$\int_\Omega u_{x_i} v\,dx = \int_\Gamma u\,v\,\nu_i ds - \int_\Omega u\,v_{x_i} dx \quad \forall\, u, v \in C^1(\overline{\Omega}).$$

This is often called *Gauss's formula* or the *divergence theorem*; and for planar regions Ω, it is also called *Green's formula*. This formula can be extended to functions from certain Sobolev spaces so that the smoothness of the functions is exactly enough for the integrals to be well-defined in the sense of Lebesgue.

Proposition 7.6.1 *Assume $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain with boundary Γ. Then*

$$\int_\Omega u_{x_i} v\,dx = \int_\Gamma u\,v\,\nu_i ds - \int_\Omega u\,v_{x_i} dx \quad \forall\, u, v \in H^1(\Omega). \tag{7.6.1}$$

Proof. Since $C^1(\overline{\Omega})$ is dense in $H^1(\Omega)$, we have sequences $\{u_n\}, \{v_n\} \subset C^1(\overline{\Omega})$, such that

$$\|u_n - u\|_{H^1(\Omega)} \to 0 \text{ and } \|v_n - v\|_{H^1(\Omega)} \to 0 \quad \text{as } n \to \infty.$$

Since $u_n, v_n \in C^1(\overline{\Omega})$, we have

$$\int_\Omega (u_n)_{x_i} v_n \, dx = \int_\Gamma u_n \, v_n \, \nu_i \, ds - \int_\Omega u_n \, (v_n)_{x_i} dx. \tag{7.6.2}$$

We take the limit as $n \to \infty$ in (7.6.2). Let us bound

$$\begin{aligned}
&\left| \int_\Omega u_{x_i} v \, dx - \int_\Omega (u_n)_{x_i} v_n \, dx \right| \\
&\qquad \le \int_\Omega |(u - u_n)_{x_i}| \, |v| \, dx + \int_\Omega |(u_n)_{x_i}| \, |v - v_n| \, dx \\
&\qquad \le \|u - u_n\|_{H^1(\Omega)} \|v\|_{L^2(\Omega)} + \|u_n\|_{H^1(\Omega)} \|v - v_n\|_{L^2(\Omega)}.
\end{aligned}$$

Since the sequences $\{u_n\}$ and $\{v_n\}$ are convergent in $H^1(\Omega)$, the quantities $\|u_n\|_{H^1(\Omega)}$ are uniformly bounded. Hence,

$$\int_\Omega u_{x_i} v \, dx - \int_\Omega (u_n)_{x_i} v_n \, dx \to 0 \quad \text{as } n \to \infty.$$

Similarly,

$$\int_\Omega u \, v_{x_i} dx - \int_\Omega u_n (v_n)_{x_i} dx \to 0 \quad \text{as } n \to \infty.$$

With regard to the boundary integral terms, we need to use the continuity of the trace operator from $H^1(\Omega)$ to $L^2(\Gamma)$. From this, we see that

$$\|u_n - u\|_{L^2(\Gamma)} \le c \, \|u_n - u\|_{H^1(\Omega)} \to 0 \quad \text{as } n \to \infty,$$

and similarly,

$$\|v_n - v\|_{L^2(\Gamma)} \to 0 \quad \text{as } n \to \infty.$$

Then use the argument technique above,

$$\int_\Gamma u \, v \, \nu_i ds - \int_\Gamma u_n v_n \, \nu_i ds \to 0 \quad \text{as } n \to \infty.$$

Taking the limit $n \to \infty$ in (7.6.2) we obtain (7.6.1). □

The above proof technique is called a "density argument". Roughly speaking, a classical integral relation can often be extended to functions from certain Sobolev spaces, as long as all the expressions in the integral relation make sense. The formula (7.6.1) suffices in studying linear second-order boundary value problems. For analyzing nonlinear problems, it is beneficial to extend the formula (7.6.1) even further. Indeed we have

$$\int_\Omega u_{x_i} v \, dx = \int_\Gamma u \, v \, \nu_i ds - \int_\Omega u \, v_{x_i} dx \quad \forall \, u \in W^{1,p}(\Omega), \; v \in W^{1,p^*}(\Omega), \tag{7.6.3}$$

where $p \in (1,\infty)$, and $p^* \in (1,\infty)$ is the conjugate exponent defined through the relation $1/p + 1/p^* = 1$.

Various other useful formulas can be derived from (7.6.1). One such formula is

$$\int_\Omega \Delta u\, v\, dx = \int_\Gamma \frac{\partial u}{\partial \nu} v\, ds - \int_\Omega \nabla u \cdot \nabla v\, dx \quad \forall\, u \in H^2(\Omega),\ v \in H^1(\Omega). \tag{7.6.4}$$

Here,

$$\Delta : u \mapsto \Delta u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$$

is the Laplacian operator,

$$\nabla u = (u_{x_1}, \ldots, u_{x_d})^T$$

is the gradient of u, and

$$\frac{\partial u}{\partial \nu} = \nabla u \cdot \boldsymbol{\nu}$$

is the outward normal derivative.

Another useful formula derived from (7.6.2) is

$$\int_\Omega (\operatorname{div} \boldsymbol{u})\, v\, dx = \int_\Gamma u_\nu v\, ds - \int_\Omega \boldsymbol{u} \cdot \nabla v\, dx \quad \forall\, \boldsymbol{u} \in H^1(\Omega)^d,\ v \in H^1(\Omega). \tag{7.6.5}$$

Here $\boldsymbol{u} = (u_1, \ldots, u_d)^T$ is a vector-valued function,

$$\operatorname{div} \boldsymbol{u} = \sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i}$$

is the divergence of $\boldsymbol{u}$, and $u_\nu = \boldsymbol{u} \cdot \boldsymbol{\nu}$ is the normal component of $\boldsymbol{u}$ on Γ.

Exercise 7.6.1 Use the density argument to prove the formula (7.6.3).

Exercise 7.6.2 Prove the formulas (7.6.4) and (7.6.5) by using (7.6.1).

Exercise 7.6.3 In the study of some thin plate deformation problems, the following integration by parts formula is useful: For any $v \in H^3(\Omega)$ and any $w \in H^2(\Omega)$,

$$\int_\Omega (2\,\partial_{12} v\, \partial_{12} w - \partial_{11} v\, \partial_{22} w - \partial_{22} v\, \partial_{11} w)\, dx = \int_\Gamma (-\partial_{\tau\tau} v\, \partial_\nu w + \partial_{\nu\tau} v\, \partial_\tau w)\, ds,$$

where Ω is a two-dimensional Lipschitz domain, $(\nu_1, \nu_2)^T$ is the unit outward normal vector on $\partial\Omega$, $(\tau_1, \tau_2)^T = (-\nu_2, \nu_1)$ is the unit tangential vector along Γ,

and

$$\partial_\nu v = \sum_{i=1}^{2} \nu_i \partial_i v, \quad \partial_\tau v = \sum_{i=1}^{2} \tau_i \partial_i v,$$
$$\partial_{\nu\tau} v = \sum_{i,j=1}^{2} \nu_i \tau_j \partial_{ij} v, \quad \partial_{\tau\tau} v = \sum_{i,j=1}^{2} \tau_i \tau_j \partial_{ij} v.$$

These derivatives as well as the normal and tangential vectors are defined a.e. on the boundary of the Lipschitz domain.

Prove the above integration by parts formula.

Suggestion for Further Reading.

Adams [1] and Lions and Magenes [158] provides a comprehensive treatment of basic aspects of Sobolev spaces, including proofs of various results. Many references on modern PDEs contain an introduction to the theory of Sobolev spaces, e.g., Evans [78], McOwen [168], Wloka [233]. For a detailed study of Sobolev spaces over non-smooth domains (domains with corners or edges), see Grisvard [98].

Sobolev spaces of any real order (i.e., $W^{s,p}(\Omega)$ with $s \in \mathbb{R}$) can be studied in the framework of interpolation spaces. Several methods are possible to develop a theory of interpolation spaces; see Triebel [225]. A relatively easily accessible reference on the topic is Bergh and Löfström [34].

8

Weak Formulations of Elliptic Boundary Value Problems

In this chapter, we consider weak formulations of some elliptic boundary value problems and study the well-posedness of the variational problems. We begin with a derivation of a weak formulation of the homogeneous Dirichlet boundary value problem for the Poisson equation. In the abstract form, a weak formulation can be viewed as an operator equation. In the second section, we provide some general results on existence and uniqueness for linear operator equations. In the third section, we present and discuss the well-known Lax-Milgram Lemma, which is applied, in the section following, to the study of well-posedness of variational formulations for various linear elliptic boundary value problems. We also apply the Lax-Milgram Lemma in studying a boundary value problem in linearized elasticity; this is done in Section 8.5. The framework in the Lax-Milgram Lemma is suitable for the development of the Galerkin method for numerically solving linear elliptic boundary value problems. In Section 8.6, we provide a brief discussion of two different weak formulations: the mixed formulation and the dual formulation. For the development of Petrov-Galerkin method, where the trial function space and the test function space are different, we discuss a generalization of Lax-Milgram Lemma in Section 8.7. Most of the chapter is concerned with boundary value problems with linear differential operators. In the last section, we analyze a nonlinear elliptic boundary value problem.

Recall that we use Ω to denote an open bounded set in $\mathbb{R}^d$, and we assume the boundary $\Gamma = \partial\Omega$ is Lipschitz continuous. Occasionally, we need to further assume Ω to be connected, and we state this assumption explicitly when it is needed.

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_8,
© Springer Science + Business Media, LLC 2009

8.1 A model boundary value problem

To begin, we use the following model boundary value problem as an illustrative example:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases} \tag{8.1.1}$$

The differential equation in (8.1.1) is called the Poisson equation. The Poisson equation can be used to describe many physical processes, e.g., steady state heat conduction, electrostatics, deformation of a thin elastic membrane (see [192]). We introduce a weak solution of the problem and discuss its relation to a classical solution of the problem.

A classical solution of the problem (8.1.1) is a smooth function $u \in C^2(\Omega) \cap C(\overline{\Omega})$ which satisfies the differential equation $(8.1.1)_1$ and the boundary condition $(8.1.1)_2$ pointwise. Necessarily we have to assume $f \in C(\Omega)$, but this condition, or even the stronger condition $f \in C(\overline{\Omega})$, does not guarantee the existence of a classical solution of the problem (see [98]). One purpose of the introduction of the weak formulation is to remove the high smoothness requirement on the solution and as a result it is easier to have the existence of a (weak) solution.

To derive the weak formulation corresponding to (8.1.1), we temporarily assume it has a classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$. We multiply the differential equation $(8.1.1)_1$ by an arbitrary function $v \in C_0^\infty(\Omega)$ (so-called smooth test functions), and integrate the relation on Ω,

$$-\int_\Omega \Delta u \, v \, dx = \int_\Omega f \, v \, dx.$$

An integration by parts for the integral on the left side yields (recall that $v = 0$ on Γ)

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx. \tag{8.1.2}$$

This relation was proved under the assumptions $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and $v \in C_0^\infty(\Omega)$. However, for all the terms in the relation (8.1.2) to make sense, we only need to require $u, v \in H^1(\Omega)$, assuming $f \in L^2(\Omega)$. Recalling the homogeneous Dirichlet boundary condition $(8.1.1)_2$, we thus seek a solution $u \in H_0^1(\Omega)$ satisfying the relation (8.1.2) for any $v \in C_0^\infty(\Omega)$. Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, the relation (8.1.2) is then valid for any $v \in H_0^1(\Omega)$. Therefore, we settle down with the following weak formulation of the boundary value problem (8.1.1):

$$u \in H_0^1(\Omega), \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx \quad \forall \, v \in H_0^1(\Omega). \tag{8.1.3}$$

Actually, we can even weaken the assumption $f \in L^2(\Omega)$. It is sufficient to assume $f \in H^{-1}(\Omega) = (H_0^1(\Omega))'$, as long as we interpret the integral

$\int_\Omega f\,v\,dx$ as the duality pairing $\langle f, v\rangle$ between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. We adopt the convention of using $\int_\Omega f\,v\,dx$ for $\langle f, v\rangle$ when $f \in H^{-1}(\Omega)$ and $v \in H_0^1(\Omega)$.

We have shown that if u is a classical solution of (8.1.1), then it is also a solution of the weak formulation (8.1.3). Conversely, suppose u is a weak solution with the additional regularity $u \in C^2(\Omega) \cap C(\overline{\Omega})$, and $f \in C(\Omega)$. Then for any $v \in C_0^\infty(\Omega) \subset H_0^1(\Omega)$, from (8.1.3) we obtain

$$\int_\Omega (-\Delta u - f)\,v\,dx = 0.$$

Then we must have $-\Delta u = f$ in Ω, i.e., the differential equation $(8.1.1)_1$ is satisfied. Also u satisfies the homogeneous Dirichlet boundary condition pointwisely.

Thus we have shown that the boundary value problem (8.1.1) and the variational problem (8.1.3) are formally equivalent. In case the weak solution u does not have the regularity $u \in C^2(\Omega) \cap C(\overline{\Omega})$, we will say u formally solves the boundary value problem (8.1.1).

We let $V = H_0^1(\Omega)$, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ the bilinear form defined by

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v\,dx \quad \text{for } u, v \in V,$$

and $\ell : V \to \mathbb{R}$ the linear functional defined by

$$\ell(v) = \int_\Omega f v\,dx \quad \text{for } v \in V.$$

Then the weak formulation of the problem (8.1.1) is to find $u \in V$ such that

$$a(u, v) = \ell(v) \quad \forall\, v \in V. \tag{8.1.4}$$

We define a differential operator A associated with the boundary value problem (8.1.1) by

$$A : H_0^1(\Omega) \to H^{-1}(\Omega), \quad \langle Au, v\rangle = a(u, v) \quad \forall\, u, v \in H_0^1(\Omega).$$

Here, $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. Then the problem (8.1.4) can be viewed as a linear operator equation

$$Au = \ell \quad \text{in } H^{-1}(\Omega).$$

A formulation of the type (8.1.1) in the form of a partial differential equation and a set of boundary conditions is referred to as a classical formulation of a boundary value problem, whereas a formulation of the type (8.1.4) is known as a weak formulation. One advantage of weak formulations over classical formulations is that questions related to existence and uniqueness of solutions can be answered more satisfactorily. Another advantage is that weak formulations naturally lead to the development of Galerkin type numerical methods (see Section 9.1).

8.2 Some general results on existence and uniqueness

We first present some general ideas and results on existence and uniqueness for a linear operator equation of the form

$$u \in V, \quad Lu = f, \tag{8.2.1}$$

where $L : \mathcal{D}(L) \subset V \to W$, V and W are Hilbert spaces, and $f \in W$. Notice that the solvability of the equation is equivalent to the property $\mathcal{R}(L) = W$, whereas the uniqueness of a solution is equivalent to the property $\mathcal{N}(L) = \{0\}$.

A very basic existence result is the following theorem.

Theorem 8.2.1 *Let V and W be Hilbert spaces, $L : \mathcal{D}(L) \subset V \to W$ a linear operator. Then $\mathcal{R}(L) = W$ if and only if $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$.*

Proof. If $\mathcal{R}(L) = W$, then obviously $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$.

Now assume $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$, but $\mathcal{R}(L) \neq W$. Then $\mathcal{R}(L)$ is a closed subspace of W. Let $w \in W \backslash \mathcal{R}(L)$. By Theorem 3.3.7, the compact set $\{w\}$ and the closed convex set $\mathcal{R}(L)$ can be strictly separated by a closed hyperplane, i.e., there exists a $w^* \in W'$ such that $\langle w^*, w\rangle > 0$ and $\langle w^*, Lv\rangle \leq 0$ for all $v \in \mathcal{D}(L)$. Since L is a linear operator, $\mathcal{D}(L)$ is a subspace of V. Hence, $\langle w^*, Lv\rangle = 0$ for all $v \in \mathcal{D}(L)$. Therefore, $0 \neq w^* \in \mathcal{R}(L)^\perp$. This is a contradiction. □

Let us see under what conditions $\mathcal{R}(L)$ is closed. We first introduce an important generalization of the notion of continuity.

Definition 8.2.2 *Let V and W be Banach spaces. An operator $T : \mathcal{D}(T) \subset V \to W$ is said to be a* closed operator *if for any sequence $\{v_n\} \subset \mathcal{D}(T)$, $v_n \to v$ and $T(v_n) \to w$ imply $v \in \mathcal{D}(T)$ and $w = T(v)$.*

We notice that a continuous operator is closed. The next example shows that a closed operator is not necessarily continuous.

Example 8.2.3 Let us consider a linear differential operator, $Lv = -\Delta v$. This operator is not continuous from $L^2(\Omega)$ to $L^2(\Omega)$. Nevertheless, L is a closed operator on $L^2(\Omega)$. To see this, let $\{v_n\}$ be a sequence converging to v in $L^2(\Omega)$ such that the sequence $\{-\Delta v_n\}$ converges to w in $L^2(\Omega)$. In the relation

$$\int_\Omega (-\Delta v_n)\, \phi \, dx = -\int_\Omega v_n \Delta \phi \, dx \quad \forall\, \phi \in C_0^\infty(\Omega)$$

we take the limit $n \to \infty$ to obtain

$$\int_\Omega w\,\phi\,dx = -\int_\Omega v\,\Delta\phi\,dx \quad \forall\,\phi \in C_0^\infty(\Omega).$$

Therefore $w = -\Delta v \in L^2(\Omega)$, and the operator L is closed. □

Theorem 8.2.4 *Let V and W be Hilbert spaces, $L : \mathcal{D}(L) \subset V \to W$ a linear closed operator. Assume for some constant $c > 0$, the following a priori estimate holds:*

$$\|Lv\|_W \geq c\,\|v\|_V \quad \forall\, v \in \mathcal{D}(L), \tag{8.2.2}$$

which is usually called a stability estimate. Also assume $\mathcal{R}(L)^\perp = \{0\}$. Then for each $f \in W$, the equation (8.2.1) *has a unique solution.*

Proof. Let us verify that $\mathcal{R}(L)$ is closed. Let $\{f_n\}$ be a sequence in $\mathcal{R}(L)$, converging to f. Then there is a sequence $\{v_n\} \subset \mathcal{D}(L)$ with $f_n = Lv_n$. By (8.2.2),

$$\|v_n - v_m\| \leq c\,\|f_n - f_m\|.$$

Thus $\{v_n\}$ is a Cauchy sequence in V. Since V is a Hilbert space, the sequence $\{v_n\}$ converges: $v_n \to v \in V$. Now L is assumed to be closed, we conclude that $v \in \mathcal{D}(L)$ and $f = Lv \in \mathcal{R}(L)$. So we can invoke Theorem 8.2.1 to obtain the existence of a solution. The uniqueness of the solution follows from the stability estimate (8.2.2). □

Noticing that a continuous operator is closed, we can replace the closedness of the operator by the continuity of the operator in Theorem 8.2.4.

Example 8.2.5 Let V be a Hilbert space, $L \in \mathcal{L}(V, V')$ be strongly monotone, i.e., for some constant $c > 0$,

$$\langle Lv, v\rangle \geq c\,\|v\|_V^2 \quad \forall\, v \in V.$$

Then (8.2.2) holds because from the monotonicity,

$$\|Lv\|_{V'}\,\|v\|_V \geq c\,\|v\|_V^2$$

which implies

$$\|Lv\|_{V'} \geq c\,\|v\|_V.$$

Also $\mathcal{R}(L)^\perp = \{0\}$, since from $v \perp \mathcal{R}(L)$ we have

$$c\,\|v\|_V^2 \leq \langle Lv, v\rangle = 0,$$

and hence $v = 0$. Therefore from Theorem 8.2.4, under the stated assumptions, for any $f \in V'$, there is a unique solution $u \in V$ to the equation $Lu = f$ in V'. □

Example 8.2.6 As a concrete example, we consider the weak formulation of the model elliptic boundary value problem (8.1.1). Here, $\Omega \subset \mathbb{R}^d$ is an open bounded set with a Lipschitz boundary $\partial\Omega$, $V = H_0^1(\Omega)$ with the norm $\|v\|_V = |v|_{H^1(\Omega)}$, and $V' = H^{-1}(\Omega)$. Given $f \in H^{-1}(\Omega)$, consider the problem

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{8.2.3}$$

We define the operator $L : V \to V'$ by

$$\langle Lu, v\rangle = \int_\Omega \nabla u \cdot \nabla v \, dx, \quad u, v \in V.$$

Then L is linear, continuous, and strongly monotone; indeed, we have

$$\|L\| = 1$$

and

$$\langle Lv, v\rangle = \|v\|_V^2 \quad \forall\, v \in V.$$

Thus from Example 8.2.5, for any $f \in H^{-1}(\Omega)$, there is a unique $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \langle f, v\rangle \quad \forall\, v \in V,$$

i.e., the boundary value problem (8.2.3) has a unique weak solution. □

It is possible to extend the existence results presented in Theorems 8.2.1 and 8.2.4 to linear operator equations on Banach spaces. Let V and W be Banach spaces, with duals V' and W'. Let $L : \mathcal{D}(L) \subset V \to W$ be a densely defined linear operator, i.e., L is a linear operator and $\mathcal{D}(L)$ is a dense subspace of V. Because $\mathcal{D}(L)$ is dense in V, one can define the dual operator $L^* : \mathcal{D}(L^*) \subset W' \to V'$ by

$$\langle L^* w^*, v\rangle = \langle w^*, Lv\rangle \quad \forall\, v \in \mathcal{D}(L),\ w^* \in W'.$$

We then define

$$\mathcal{N}(L)^\perp = \{v^* \in V' \mid \langle v^*, v\rangle = 0 \ \ \forall\, v \in \mathcal{N}(L)\},$$
$$\mathcal{N}(L^*)^\perp = \{w \in W \mid \langle w^*, w\rangle = 0 \ \ \forall\, w^* \in \mathcal{N}(L^*)\}.$$

The most important theorem on dual operators in Banach spaces is the following closed range theorem of Banach (see e.g., [250, p. 210]).

Theorem 8.2.7 *Assume V and W are Banach spaces, $L : \mathcal{D}(L) \subset V \to W$ is a densely defined linear closed operator. Then the following four statements are equivalent.*
(a) *$\mathcal{R}(L)$ is closed in W.*
(b) *$\mathcal{R}(L) = \mathcal{N}(L^*)^\perp$.*
(c) *$\mathcal{R}(L^*)$ is closed in V'.*
(d) *$\mathcal{R}(L^*) = \mathcal{N}(L)^\perp$.*

In particular, this theorem implies the abstract Fredholm alternative result: If $\mathcal{R}(L)$ is closed, then $\mathcal{R}(L) = \mathcal{N}(L^*)^\perp$, i.e., the equation $Lu = f$ has a solution $u \in \mathcal{D}(L)$ if and only if $\langle w^*, f\rangle = 0$ for any $w^* \in W'$ with $L^* w^* = 0$. The closedness of $\mathcal{R}(L)$ follows from the stability estimate

$$\|Lv\| \geq c\,\|v\| \quad \forall\, v \in \mathcal{D}(L),$$

as we have seen in the proof of Theorem 8.2.4.

Now we consider the issue of uniqueness of a solution to a nonlinear operator equation. We have the following general result.

Theorem 8.2.8 *Assume V and W are Banach spaces, $T : \mathcal{D}(T) \subset V \to W$. Then for any $w \in W$, there exists at most one solution $u \in V$ of the equation $T(u) = w$, if one of the following conditions is satisfied.*
(a) *Stability: for some constant $c > 0$,*

$$\|T(u) - T(v)\| \geq c\,\|u - v\| \quad \forall\, u, v \in \mathcal{D}(T).$$

(b) *Contractivity of $T - I$:*

$$\|(T(u) - u) - (T(v) - v)\| < \|u - v\| \quad \forall\, u, v \in \mathcal{D}(T),\ u \neq v.$$

Proof. (a) Assume both u_1 and u_2 are solutions. Then $T(u_1) = T(u_2) = w$. Apply the stability condition,

$$c\,\|u_1 - u_2\| \leq \|T(u_1) - T(u_2)\| = 0.$$

Therefore, $u_1 = u_2$.

(b) Suppose there are two solutions $u_1 \neq u_2$. Then from the contractivity condition, we have

$$\|u_1 - u_2\| > \|(T(u_1) - u_1) - (T(u_2) - u_2)\| = \|u_1 - u_2\|.$$

This is a contradiction. □

We remark that the result of Theorem 8.2.8 certainly holds in the special case of Hilbert spaces V and W, and when $T = L : \mathcal{D}(L) \subset V \to W$ is a linear operator. In the case of a linear operator, the stability condition reduces to the estimate (8.2.2).

Exercise 8.2.1 Consider a linear system on $\mathbb{R}^d$: $A\,\boldsymbol{x} = \boldsymbol{b}$, where $A \in \mathbb{R}^{d\times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$. Recall the well-known result that for such a linear system, existence and uniqueness are equivalent. Apply Theorem 8.2.8 to find sufficient conditions on A which guarantee the unique solvability of the linear system for any given $\boldsymbol{b} \in \mathbb{R}^d$.

8.3 The Lax-Milgram Lemma

The Lax-Milgram Lemma is employed frequently in the study of linear elliptic boundary value problems of the form (8.1.4). For a real Banach space V, let us first explore the relation between a linear operator $A : V \to V'$ and a bilinear form $a : V \times V \to \mathbb{R}$ related by

$$\langle Au, v\rangle = a(u, v) \quad \forall\, u, v \in V. \tag{8.3.1}$$

The bilinear form $a(\cdot, \cdot)$ is continuous if and only if there exists $M > 0$ such that

$$|a(u, v)| \le M \|u\| \|v\| \quad \forall\, u, v \in V.$$

Theorem 8.3.1 *There exists a one-to-one correspondence between linear continuous operators $A : V \to V'$ and continuous bilinear forms $a : V \times V \to \mathbb{R}$, given by the formula* (8.3.1).

Proof. If $A \in \mathcal{L}(V, V')$, then $a : V \times V \to \mathbb{R}$ defined in (8.3.1) is bilinear and bounded:

$$|a(u, v)| \le \|Au\| \|v\| \le \|A\| \|u\| \|v\| \quad \forall\, u, v \in V.$$

Conversely, let $a(\cdot, \cdot)$ be given as a continuous bilinear form on V. For any fixed $u \in V$, the map $v \mapsto a(u, v)$ defines a linear continuous operator on V. Thus, there is an element $Au \in V'$ such that (8.3.1) holds. From the bilinearity of $a(\cdot, \cdot)$, we obtain the linearity of A. From the boundedness of $a(\cdot, \cdot)$, we obtain the boundedness of A. □

With a linear operator A and a bilinear form a related through (8.3.1), many properties of the linear operator A can be defined through those of the bilinear form a, or vice versa. Some examples are (assuming V is a real Hilbert space):

- a is bounded ($a(u, v) \le M \|u\| \|v\| \ \forall\, u, v \in V$) if and only if A is bounded ($\|Av\| \le M \|v\| \ \forall\, v \in V$).
- a is positive ($a(v, v) \ge 0 \ \forall\, v \in V$) if and only if A is positive ($\langle Av, v\rangle \ge 0 \ \forall\, v \in V$).
- a is strictly positive ($a(v, v) > 0 \ \forall\, 0 \ne v \in V$) if and only if A is strictly positive ($\langle Av, v\rangle > 0 \ \forall\, 0 \ne v \in V$).
- a is strongly positive or V-elliptic ($a(v, v) \ge \alpha \|v\|^2 \ \forall\, v \in V$, for some constant $\alpha > 0$) if and only if A is strongly positive ($\langle Av, v\rangle \ge \alpha \|v\|^2 \ \forall\, v \in V$).
- a is symmetric ($a(u, v) = a(v, u) \ \forall\, u, v \in V$) if and only if A is symmetric ($\langle Au, v\rangle = \langle Av, u\rangle \ \forall\, u, v \in V$).

We now recall the following minimization principle from Chapter 3.

Theorem 8.3.2 *Assume K is a nonempty, closed, convex subset of the Hilbert space V, $\ell \in V'$. Let*

$$E(v) = \frac{1}{2}\|v\|^2 - \ell(v), \quad v \in V.$$

Then there exists a unique $u \in K$ such that

$$E(u) = \inf_{v \in K} E(v).$$

The minimizer u is uniquely characterized by the inequality

$$u \in K, \quad (u, v-u) \geq \ell(v-u) \quad \forall\, v \in K.$$

If additionally, K is a subspace of V, then u is equivalently defined by

$$u \in K, \quad (u, v) = \ell(v) \quad \forall\, v \in K.$$

Let us apply this result to get the Lax-Milgram Lemma in the case the bilinear form is symmetric.

Theorem 8.3.3 *Assume K is a nonempty, closed, convex subset of the Hilbert space V, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is bilinear, symmetric, bounded and V-elliptic, and $\ell \in V'$. Let*

$$E(v) = \frac{1}{2}\, a(v, v) - \ell(v), \quad v \in V.$$

Then there exists a unique $u \in K$ such that

$$E(u) = \inf_{v \in K} E(v), \tag{8.3.2}$$

which is also the unique solution of the variational inequality

$$u \in K, \quad a(u, v-u) \geq \ell(v-u) \quad \forall\, v \in K, \tag{8.3.3}$$

or

$$u \in K, \quad a(u, v) = \ell(v) \quad \forall\, v \in K \tag{8.3.4}$$

in the special case where K is a subspace.

Proof. By the assumptions,

$$(u, v)_a = a(u, v), \quad u, v \in V$$

defines an inner product on V with the induced norm

$$\|v\|_a = \sqrt{a(v, v)}$$

which is equivalent to the original norm:

$$c_1\|v\| \le \|v\|_a \le c_2\|v\| \quad \forall\, v \in V$$

for some constants $0 < c_1 \le c_2 < \infty$. Also notice that ℓ is continuous with respect to the original norm if and only if it is continuous with respect to the norm $\|\cdot\|_a$. Now

$$E(v) = \frac{1}{2}\,\|v\|_a^2 - \ell(v)$$

and we can apply the results of Theorem 8.3.2. □

In case the bilinear form $a(\cdot,\cdot)$ is not symmetric, there is no longer an associated minimization problem, yet we can still discuss the solvability of the variational equation (8.3.4) (the next theorem) or the variational inequality (8.3.3) (see Chapter 11).

Theorem 8.3.4 (Lax-Milgram Lemma) *Assume V is a Hilbert space, $a(\cdot,\cdot)$ is a bounded, V-elliptic bilinear form on V, $\ell \in V'$. Then there is a unique solution of the problem*

$$u \in V, \quad a(u,v) = \ell(v) \quad \forall\, v \in V. \tag{8.3.5}$$

Before proving the result, let us consider the simple real linear equation

$$x \in \mathbb{R}\,, \quad a\,x = \ell.$$

Its weak formulation is

$$x \in \mathbb{R}\,, \quad a\,x\,y = \ell\,y \quad \forall\, y \in \mathbb{R}\,.$$

We observe that the real linear equation has a solution if and only if $0 < a < \infty$ (we multiply the equation by (-1) to make a positive, if necessary) and $|\ell| < \infty$, i.e., if and only if the bilinear form $a(x,y) \equiv a\,x\,y$ is continuous and $\mathbb{R}$-elliptic, and the linear form $\ell(y) \equiv \ell\,y$ is bounded. Thus the assumptions made in Theorem 8.3.4 are quite natural.

Several different proofs are possible for this important result. Here we present two of them.

Proof. **[#1]** For any $\theta > 0$, the problem (8.3.5) is equivalent to

$$(u,v) = (u,v) - \theta\,[a(u,v) - \ell(v)] \quad \forall\, v \in V,$$

i.e., the fixed-point problem

$$u = P_\theta(u),$$

where $P_\theta(u) \in V$ is defined through the relation

$$(P_\theta(u), v) = (u,v) - \theta\,[a(u,v) - \ell(v)], \quad v \in V.$$

We will apply the Banach fixed-point theorem with a proper choice of θ. Let $A : V \to V'$ be the linear operator associated with the bilinear form $a(\cdot,\cdot)$; see (8.3.1). Then A is bounded and strongly positive:

$$\|Av\| \le M\,\|v\| \text{ and } \langle Av, v\rangle \ge \alpha\,\|v\|^2 \quad \forall\, v \in V.$$

Denote $\mathcal{J} : V' \to V$ the isometric dual mapping from the Riesz representation theorem. Then

$$a(u,v) = \langle Au, v\rangle = (\mathcal{J}Au, v) \quad \forall\, u, v \in V,$$

and

$$\|\mathcal{J}Au\| = \|Au\| \quad \forall\, u \in V.$$

For any $u \in V$, by Theorem 8.3.3, the problem

$$(w,v) = (u,v) - \theta\,[a(u,v) - \ell(v)] \quad \forall\, v \in V$$

has a unique solution $w = P_\theta(u)$. Let us show that for $\theta \in (0, 2\,\alpha/M^2)$, the operator P_θ is a contraction. Indeed let $u_1, u_2 \in V$, and denote $w_1 = P_\theta(u_1)$, $w_2 = P_\theta(u_2)$. Then

$$(w_1 - w_2, v) = (u_1 - u_2, v) - \theta\, a(u_1 - u_2, v) = ((I - \theta\,\mathcal{J}A)(u_1 - u_2), v),$$

i.e., $w_1 - w_2 = (I - \theta\,\mathcal{J}A)(u_1 - u_2)$. We then have

$$\begin{aligned}\|w_1 - w_2\|^2 &= \|u_1 - u_2\|^2 - 2\,\theta\,(\mathcal{J}A(u_1 - u_2), u_1 - u_2)\\ &\quad + \theta^2\|\mathcal{J}A(u_1 - u_2)\|^2\\ &= \|u_1 - u_2\|^2 - 2\,\theta\, a(u_1 - u_2, u_1 - u_2) + \theta^2\|A(u_1 - u_2)\|^2\\ &\le (1 - 2\,\theta\,\alpha + \theta^2 M^2)\,\|u_1 - u_2\|^2.\end{aligned}$$

Since $\theta \in (0, 2\,\alpha/M^2)$, we have

$$1 - 2\,\theta\,\alpha + \theta^2 M^2 < 1$$

and the mapping P_θ is a contraction. By the Banach fixed-point theorem, P_θ has a unique fixed-point $u \in V$, which is the solution of the problem (8.3.5).

[#2] The uniqueness of a solution follows from the V-ellipticity of the bilinear form. We prove the existence by applying Theorem 8.2.1. We will use the linear operator $L = \mathcal{J}A : V \to V$ constructed in the first proof. We recall that $\mathcal{R}(L) = V$ if and only if $\mathcal{R}(L)$ is closed and $\mathcal{R}(L)^\perp = \{0\}$.

To show $\mathcal{R}(L)$ is closed, we let $\{u_n\} \subset \mathcal{R}(L)$ be a sequence converging to u. Then $u_n = \mathcal{J}Aw_n$ for some $w_n \in V$. We have

$$\|u_n - u_m\| = \|\mathcal{J}A(w_n - w_m)\| = \|A(w_n - w_m)\| \ge \alpha\,\|w_n - w_m\|.$$

Hence $\{w_n\}$ is a Cauchy sequence and so has a limit $w \in V$. Then

$$\|u_n - \mathcal{J}Aw\| = \|\mathcal{J}A(w_n - w)\| = \|A(w_n - w)\| \leq M\,\|w_n - w\| \to 0.$$

Hence, $u = \mathcal{J}Aw \in \mathcal{R}(L)$ and $\mathcal{R}(L)$ is closed.

Now suppose $u \in \mathcal{R}(L)^{\perp}$. Then for any $v \in V$,

$$0 = (\mathcal{J}Av, u) = a(v, u).$$

Taking $v = u$ above, we have $a(u, u) = 0$. By the V-ellipticity of $a(\cdot, \cdot)$, we conclude $u = 0$. □

Example 8.3.5 Applying the Lax-Milgram Lemma, we conclude that the boundary value problem (8.2.3) has a unique weak solution $u \in H_0^1(\Omega)$. □

Exercise 8.3.1 Deduce the Lax-Milgram Lemma from Theorem 5.1.4.

8.4 Weak formulations of linear elliptic boundary value problems

In this section, we formulate and analyze weak formulations of some linear elliptic boundary value problems. To present the ideas clearly, we will frequently refer to boundary value problems associated with the Poisson equation

$$-\Delta u = f$$

and the Helmholtz equation

$$-\Delta u + u = f$$

for examples.

8.4.1 Problems with homogeneous Dirichlet boundary conditions

So far, we have studied the model elliptic boundary value problem corresponding to the Poisson equation with the homogeneous Dirichlet boundary condition

$$-\Delta u = f \quad \text{in } \Omega, \tag{8.4.1}$$

$$u = 0 \quad \text{in } \Gamma, \tag{8.4.2}$$

where $f \in L^2(\Omega)$ is given. The weak formulation of the problem is

$$u \in V, \quad a(u, v) = \ell(v) \quad \forall\, v \in V. \tag{8.4.3}$$

Here

$$V = H_0^1(\Omega),$$
$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx \quad \text{for } u, v \in V,$$
$$\ell(v) = \int_\Omega f v \, dx \quad \text{for } v \in V.$$

The problem (8.4.3) has a unique solution $u \in V$ by the Lax-Milgram Lemma.

Dirichlet boundary conditions are also called *essential boundary conditions* since they are explicitly required by the weak formulations.

8.4.2 Problems with non-homogeneous Dirichlet boundary conditions

Suppose that instead of (8.4.2) the boundary condition is

$$u = g \quad \text{on } \Gamma. \tag{8.4.4}$$

To derive a weak formulation, we proceed similarly as in Section 8.1. We first assume the boundary value problem given by (8.4.1) and (8.4.4) has a classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$. Multiplying the equation (8.4.1) by a test function v with certain smoothness which validates the following calculations, and integrating over Ω, we have

$$\int_\Omega (-\Delta u)\, v \, dx = \int_\Omega f \, v \, dx.$$

Integrate by parts,

$$-\int_\Gamma \frac{\partial u}{\partial \nu} \, v \, ds + \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx.$$

We now assume $v = 0$ on Γ so that the boundary integral term vanishes; the boundary integral term would otherwise be difficult to deal with under the expected regularity condition $u \in H^1(\Omega)$ on the weak solution. Thus we arrive at the relation

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx$$

if v is smooth and $v = 0$ on Γ. For each term in the above relation to make sense, we assume $f \in L^2(\Omega)$, and let $u \in H^1(\Omega)$ and $v \in H_0^1(\Omega)$. Recall that the solution u should satisfy the boundary condition $u = g$ on Γ. We observe that it is necessary to assume $g \in H^{1/2}(\Gamma)$, i.e., g is the trace on

Γ of an $H^1(\Omega)$ function. Finally, we obtain the weak formulation for the boundary value problem of (8.4.1) and (8.4.4):

$$u \in H^1(\Omega),\ u = g \text{ on } \Gamma, \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f\, v \, dx \quad \forall\, v \in H_0^1(\Omega). \tag{8.4.5}$$

For the weak formulation (8.4.5), though, we cannot apply Lax-Milgram Lemma directly since the trial function u and the test function v do not lie in the same space. There is a standard way to get rid of this problem. Since $g \in H^{1/2}(\Gamma)$ and $\gamma(H^1(\Omega)) = H^{1/2}(\Gamma)$ (recall that γ is the trace operator), we have the existence of a function $G \in H^1(\Omega)$ such that $\gamma G = g$. We remark that finding the function G in practice may be nontrivial. Thus, with

$$u = w + G,$$

the problem may be transformed into one of seeking w such that

$$w \in H_0^1(\Omega), \quad \int_\Omega \nabla w \cdot \nabla v \, dx = \int_\Omega (f\, v - \nabla G \cdot \nabla v)\, dx \quad \forall\, v \in H_0^1(\Omega). \tag{8.4.6}$$

The classical form of the boundary value problem for w is

$$\begin{aligned} -\Delta w &= f + \Delta G \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Applying the Lax-Milgram Lemma, we have a unique solution $w \in H_0^1(\Omega)$ of the problem (8.4.6). Then we set $u = w + G$ to get a solution u of the problem (8.4.5). Notice that the choice of the function G is not unique, so the uniqueness of the solution u of the problem (8.4.5) does not follow from the above argument. Nevertheless we can show the uniqueness of u by a standard approach. Assume both u_1 and u_2 are solutions of the problem (8.4.5). Then the difference $u_1 - u_2$ satisfies

$$u_1 - u_2 \in H_0^1(\Omega), \quad \int_\Omega \nabla(u_1 - u_2) \cdot \nabla v \, dx = 0 \quad \forall\, v \in H_0^1(\Omega).$$

Taking $v = u_1 - u_2$, we obtain

$$\int_\Omega |\nabla(u_1 - u_2)|^2 dx = 0.$$

Thus, $\nabla(u_1 - u_2) = 0$ a.e. in Ω, and hence $u_1 - u_2$ is a constant in each connected component of Ω. Using the boundary condition $u_1 - u_2 = 0$ a.e. on Γ, we see that $u_1 = u_2$ a.e. in Ω.

Since non-homogeneous Dirichlet boundary conditions can be rendered homogeneous in the way described above, for convenience only problems with homogeneous Dirichlet conditions will be considered later.

8.4.3 Problems with Neumann boundary conditions

Consider next the *Neumann* problem of determining u which satisfies

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega, \\ \partial u/\partial \nu = g & \text{on } \Gamma. \end{cases} \tag{8.4.7}$$

Here f and g are given functions in Ω and on Γ, respectively, and $\partial/\partial\nu$ denotes the normal derivative on Γ. Again we first derive a weak formulation. Assume $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ is a classical solution of the problem (8.4.7). Multiplying $(8.4.7)_1$ by an arbitrary test function v with certain smoothness for the following calculations to make sense, integrating over Ω and performing an integration by parts, we obtain

$$\int_\Omega (\nabla u \cdot \nabla v + uv)\, dx = \int_\Omega fv\, dx + \int_\Gamma \frac{\partial u}{\partial \nu} v\, ds.$$

Then, substitution of the Neumann boundary condition $(8.4.7)_2$ in the boundary integral term leads to the relation

$$\int_\Omega (\nabla u \cdot \nabla v + uv)\, dx = \int_\Omega fv\, dx + \int_\Gamma g\, v\, ds.$$

Assume $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$. For each term in the above relation to make sense, it is natural to choose the space $H^1(\Omega)$ for both the trial function u and the test function v. Thus, the weak formulation of the boundary value problem (8.4.7) is

$$u \in H^1(\Omega), \quad \int_\Omega (\nabla u \cdot \nabla v + uv)\ dx = \int_\Omega fv\, dx + \int_\Gamma gv\, ds \quad \forall\, v \in H^1(\Omega). \tag{8.4.8}$$

This problem has the form (8.4.3), where $V = H^1(\Omega)$, $a(\cdot,\cdot)$ and $\ell(\cdot)$ are defined by

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + uv)\, dx,$$
$$\ell(v) = \int_\Omega fv\, dx + \int_\Gamma gv\, ds.$$

Applying the Lax-Milgram Lemma, we can show that the weak formulation (8.4.8) has a unique solution $u \in H^1(\Omega)$.

Above we have shown that a classical solution $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ of the boundary value problem (8.4.7) is also the solution $u \in H^1(\Omega)$ of the weak formulation (8.4.8). Conversely, reversing the above arguments, it is readily seen that a weak solution of the problem (8.4.8) with sufficient smoothness is also the classical solution of the problem (8.4.7).

Neumann boundary conditions are also called *natural boundary conditions* since they are naturally incorporated in the weak formulations of the boundary value problems, as can be seen from (8.4.8).

It is more delicate to study the pure Neumann problem for the Poisson equation

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \partial u/\partial \nu = g & \text{on } \Gamma, \end{cases} \tag{8.4.9}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$ are given. Formally, the corresponding weak formulation is

$$u \in H^1(\Omega), \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f\, v \, dx + \int_\Gamma g\, v \, ds \quad \forall\, v \in H^1(\Omega). \tag{8.4.10}$$

A necessary condition for (8.4.10) to have a solution is

$$\int_\Omega f \, dx + \int_\Gamma g \, ds = 0, \tag{8.4.11}$$

which is derived from (8.4.10) by taking the test function $v = 1$. If (8.4.11) is not valid, the problem (8.4.9) does not have a solution. When the problem has a solution u, any function of the form $u+c$, $c \in \mathbb{R}$, is a solution. Assume $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain, i.e. Ω is open, bounded, connected, and its boundary is Lipschitz continuous. Let us show that the condition (8.4.11) is also a sufficient condition for the problem (8.4.10) to have a solution. Indeed, the problem (8.4.10) is most conveniently studied in the quotient space $V = H^1(\Omega)/\mathbb{R}$ (see Exercise 1.2.19 for the definition of a quotient space), where each element $[v] \in V$ is an equivalence class $[v] = \{v + \alpha \mid \alpha \in \mathbb{R}\}$, and any $v \in [v] \subset H^1(\Omega)$ is called a representative element of $[v]$. The following result is a special case of Theorem 7.3.17.

Lemma 8.4.1 *Assume $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain. Then over the quotient space $V = H^1(\Omega)/\mathbb{R}$, the quotient norm*

$$\|[v]\|_V \equiv \inf_{v \in [v]} \|v\|_1 = \inf_{\alpha \in \mathbb{R}} \|v + \alpha\|_1$$

is equivalent to the $H^1(\Omega)$ semi-norm $|v|_1$ for any $v \in [v]$.

It is now easy to see that the formula

$$\overline{a}([u],[v]) = \int_\Omega \nabla u \cdot \nabla v \, dx, \quad u \in [u],\ v \in [v]$$

defines a bilinear form on V, which is continuous and V-elliptic. Because of the condition (8.4.11),

$$\overline{\ell}([v]) = \int_\Omega f\, v \, dx + \int_\Gamma g\, v \, ds$$

is a well-defined linear continuous form on V. Hence, we can apply the Lax-Milgram Lemma to conclude that the problem

$$[u] \in V, \quad \overline{a}([u],[v]) = \overline{\ell}([v]) \quad \forall\, [v] \in V$$

has a unique solution $[u]$. It is easy to see that any $u \in [u]$ is a solution of (8.4.10).

Another approach to studying the Neumann boundary value problem (8.4.9) is to add a side condition, such as

$$\int_\Omega u \, dx = 0.$$

Then we introduce the space

$$V = \left\{ v \in H^1(\Omega) \;\middle|\; \int_\Omega v \, dx = 0 \right\}. \tag{8.4.12}$$

An application of Theorem 7.3.12 shows that over the space V, $|\cdot|_1$ is a norm equivalent to the norm $\|\cdot\|_1$. The bilinear form $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$ is both continuous and V-elliptic. So there is a unique solution to the problem

$$u \in V, \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx + \int_\Gamma g \, v \, ds \quad \forall \, v \in V. \tag{8.4.13}$$

The unique solution $u \in V$ of (8.4.13) can be obtained as the limit of a sequence of weak solutions of regularized boundary value problems. See Exercise 8.4.7.

8.4.4 Problems with mixed boundary conditions

It is also possible to specify different kind of boundary conditions on different portions of the boundary. One such example is

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \partial u / \partial \nu = g & \text{on } \Gamma_N, \end{cases} \tag{8.4.14}$$

where Γ_D and Γ_N form a non-overlapping decomposition of the boundary: $\Gamma = \Gamma_D \cup \Gamma_N$, Γ_D is relatively closed, Γ_N is relatively open, and $\Gamma_D \cap \Gamma_N = \emptyset$. Assume Ω is connected. The appropriate space in which to pose this problem in weak form is now

$$V = H^1_{\Gamma_D}(\Omega) \equiv \{ v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_D \}.$$

Then the weak problem is again of the form (8.4.7) with

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + u \, v) \, dx,$$
$$\ell(v) = \int_\Omega f \, v \, dx + \int_{\Gamma_N} g \, v \, ds.$$

Under suitable assumptions, say $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_N)$, we can again apply the Lax-Milgram Lemma to conclude that the weak problem has a unique solution.

8.4.5 A general linear second-order elliptic boundary value problem

The issue of existence and uniqueness of solutions to the problems just discussed may be treated in the more general framework of arbitrary linear elliptic PDEs of second order. Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain. Let the boundary $\Gamma = \Gamma_D \cup \Gamma_N$ with $\Gamma_D \cap \Gamma_N = \emptyset$, Γ_D and Γ_N being relatively closed and open subsets of Γ. Consider the boundary value problem

$$\begin{cases} -\sum_{i,j=1}^d \partial_j(a_{ij}\partial_i u) + \sum_{i=1}^d b_i \partial_i u + cu = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \sum_{i,j=1}^d a_{ij}\partial_i u\, \nu_j = g & \text{on } \Gamma_N. \end{cases} \tag{8.4.15}$$

Here $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^T$ is the unit outward normal on Γ_N.

The given functions a_{ij}, b_i, c, f, and g are assumed to satisfy the following conditions:

$$a_{ij}, b_i, c \in L^\infty(\Omega); \tag{8.4.16}$$

there exists a constant $\theta > 0$ such that

$$\sum_{i,j=1}^d a_{ij}\xi_i\xi_j \geq \theta\,|\boldsymbol{\xi}|^2 \quad \forall\, \boldsymbol{\xi} = (\xi_i) \in \mathbb{R}^d, \text{ a.e. in } \Omega; \tag{8.4.17}$$

$$f \in L^2(\Omega); \tag{8.4.18}$$

$$g \in L^2(\Gamma_N). \tag{8.4.19}$$

The weak formulation of the problem (8.4.15) is obtained again in the usual way by multiplying the differential equation in (8.4.15) by an arbitrary test function v which vanishes on Γ_D, integrating over Ω, performing an integration by parts, and applying the specified boundary conditions. As a result, we get the weak formulation (8.4.3) with

$$\begin{aligned} V &= H^1_{\Gamma_D}(\Omega), \\ a(u,v) &= \int_\Omega \left[\sum_{i,j=1}^d a_{ij}\partial_i u \partial_j v + \sum_{i=1}^d b_i(\partial_i u)\, v + c\, u\, v \right] dx, \\ \ell(v) &= \int_\Omega f\, v\, dx + \int_{\Gamma_N} g\, v\, ds. \end{aligned} \tag{8.4.20}$$

We can again apply Lax-Milgram Lemma to study the well-posedness of the boundary value problem. The space $V = H^1_{\Gamma_D}(\Omega)$ is a Hilbert space, with the standard H^1-norm. The assumptions (8.4.16)–(8.4.19) ensure that the bilinear form is bounded on V, and the linear form is bounded on V. What remains to be established is the V-ellipticity of the bilinear form.

Some sufficient conditions for the V-ellipticity of the bilinear form of the left hand side of (8.4.20) are discussed in Exercise 8.4.3.

Exercise 8.4.1 The boundary value problem

$$\begin{cases} -\Delta u + c\,u = f & \text{in } \Omega, \\ u = \text{constant} & \text{on } \Gamma, \\ \displaystyle\int_\Gamma \frac{\partial u}{\partial \nu}\,ds = \int_\Gamma g\,ds \end{cases}$$

is called an Adler problem. Derive a weak formulation, and show that the weak formulation and the boundary value problem are formally equivalent. Assume $c > 0$, $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$. Prove that the weak formulation has a unique solution.

Exercise 8.4.2 A boundary condition can involve both the unknown function and its normal derivative; such a boundary condition is called the *third boundary condition* or *Robin boundary condition* for second-order differential equations. Consider the boundary value problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} + a\,u &= g \quad \text{on } \Gamma. \end{aligned}$$

Derive a weak formulation of the Robin boundary value problem for the Poisson equation. Find conditions on the given data for the existence and uniqueness of a solution to the weak formulation; prove your assertion.

Exercise 8.4.3 Assume $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain. Show that the bilinear form defined in (8.4.20) is V-elliptic with $V = H^1_{\Gamma_D}(\Omega)$, if (8.4.16)–(8.4.19) hold and one of the following three conditions is satisfied, with $\boldsymbol{b} = (b_1, \ldots, b_d)^T$ and θ the ellipticity constant in (8.4.17):

$$c \geq c_0 > 0,\ |\boldsymbol{b}| \leq B \text{ a.e. in } \Omega, \text{ and } B^2 < 4\,\theta\,c_0,$$

or

$$\boldsymbol{b} \cdot \boldsymbol{\nu} \geq 0 \text{ a.e. on } \Gamma_N, \text{ and } c - \frac{1}{2}\operatorname{div} \boldsymbol{b} \geq c_0 > 0 \text{ a.e. in } \Omega,$$

or

$$\operatorname{meas}(\Gamma_D) > 0,\ \boldsymbol{b} = \boldsymbol{0}, \text{ and } \inf_\Omega c > -\theta/\bar{c},$$

where $\bar{c}$ is the best constant in the Poincaré inequality

$$\int_\Omega v^2 dx \leq \bar{c} \int_\Omega |\nabla v|^2 dx \quad \forall\, v \in H^1_{\Gamma_D}(\Omega).$$

This best constant can be computed by solving a linear elliptic eigenvalue problem: $\bar{c} = 1/\lambda_1$, with $\lambda_1 > 0$ the smallest eigenvalue of the eigenvalue problem

$$\begin{cases} -\Delta u = \lambda\,u & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \Gamma_N. \end{cases}$$

A special and important case is that corresponding to $\boldsymbol{b} = \boldsymbol{0}$; in this case the bilinear form is symmetric, and V-ellipticity is assured if

$$c \geq c_0 > 0.$$

Exercise 8.4.4 It is not always necessary to assume Ω to be connected. Let $\Omega \subset \mathbb{R}^d$ be open, bounded with a Lipschitz boundary, and let us consider the boundary value problem 8.4.15 with $\Gamma_D = \Gamma$ and $\Gamma_N = \emptyset$ (i.e. a pure Dirichlet boundary value problem). Keep the assumptions (8.4.16)–(8.4.18). Show that the boundary value problem has a unique solution if one of the following three conditions is satisfied:

$$c \geq c_0 > 0, \ |\boldsymbol{b}| \leq B \text{ a.e. in } \Omega, \text{ and } B^2 < 4\,\theta\, c_0,$$

or

$$c - \frac{1}{2}\operatorname{div} \boldsymbol{b} \geq c_0 > 0 \text{ a.e. in } \Omega,$$

or

$$\boldsymbol{b} = \boldsymbol{0} \text{ and } \inf_{\Omega} c > -\theta/\bar{c}$$

where $\bar{c}$ is the best constant in the Poincaré inequality

$$\int_\Omega v^2 dx \leq \bar{c} \int_\Omega |\nabla v|^2 dx \quad \forall\, v \in H_0^1(\Omega).$$

This best constant can be computed by solving a linear elliptic eigenvalue problem: $\bar{c} = 1/\lambda_1$, with $\lambda_1 > 0$ the smallest eigenvalue of the eigenvalue problem

$$\begin{cases} -\Delta u = \lambda\, u & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases}$$

Exercise 8.4.5 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain, $f \in L^2(\Omega)$, $b_1, \cdots, b_d \in \mathbb{R}$, $c_0 \in L^\infty(\Omega)$, and $c_0 \geq 0$ a.e. in Ω. Consider the boundary value problem

$$\begin{aligned} -\Delta u + b_1 \frac{\partial u}{\partial x_1} + \cdots + b_d \frac{\partial u}{\partial x_d} + c_0 u = f \quad & \text{in } \Omega, \\ u = 0 \quad & \text{on } \Gamma. \end{aligned}$$

Give a weak formulation, study its solution existence and uniqueness, and explore the availability of an equivalent minimization problem.

Exercise 8.4.6 Assume Ω is a Lipschitz domain in $\mathbb{R}^d$. Recall the definition (7.3.2) of the norm $\|g\|_{H^{1/2}(\Gamma)}$. Show that $\|g\|_{H^{1/2}(\Gamma)} = \|u\|_{H^1(\Omega)}$ where $u \in H^1(\Omega)$ satisfies

$$\gamma u = g \text{ on } \Gamma, \quad \int_\Omega (\nabla u \cdot \nabla v + u\, v)\, dx = 0 \quad \forall\, v \in H_0^1(\Omega).$$

In other words, u is a weak solution of the boundary value problem

$$\begin{aligned} -\Delta u + u = 0 \quad & \text{in } \Omega, \\ u = g \quad & \text{on } \Gamma. \end{aligned}$$

Hint: By the definition, there exists a sequence $\{u_n\} \subset H^1(\Omega)$ with $u_n = g$ on Γ such that

$$\|u_n\|_{H^1(\Omega)} \to \|g\|_{H^{1/2}(\Gamma)} \quad \text{as } n \to \infty.$$

Show that there is a subsequence $\{u_{n'}\}$ and an element $u \in H^1(\Omega)$ such that

$$u_{n'} \rightharpoonup u \text{ in } H^1(\Omega), \quad u_{n'} \to u \text{ in } L^2(\Gamma).$$

Then deduce that $u \in H^1(\Omega)$ satisfies $u = g$ on Γ, and

$$\|u\|_{H^1(\Omega)} = \inf\{\|v\|_{H^1(\Omega)} \mid v \in H^1(\Omega),\ \gamma v = g\}.$$

Finally, notice that for fixed $v \in H_0^1(\Omega)$, the real variable function $\|u + tv\|^2_{H^1(\Omega)}$ has its minimum at $t = 0$.

Exercise 8.4.7 In this exercise, we discuss how to apply the so-called Tikhonov regularization method to approximate the solution u of the problem (8.4.13). For $\varepsilon > 0$, consider the problem

$$\begin{aligned} -\Delta u_\varepsilon + \varepsilon\, u_\varepsilon &= f \quad \text{in } \Omega, \\ \frac{\partial u_\varepsilon}{\partial \nu} &= g \quad \text{on } \Gamma, \end{aligned}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$ are given functions satisfying (8.4.11). The weak formulation is

$$u_\varepsilon \in H^1(\Omega), \quad \int_\Omega (\nabla u_\varepsilon \cdot \nabla v + \varepsilon\, u_\varepsilon v)\, dx = \int_\Omega f\, v\, dx + \int_\Gamma g\, v\, ds \quad \forall\, v \in H^1(\Omega). \tag{8.4.21}$$

Carry out the following steps.
(a) For any $\varepsilon > 0$, show (8.4.21) has a unique solution.
(b) Show that $\int_\Omega u_\varepsilon(\boldsymbol{x})\, dx = 0$.
(c) Take $v = u_\varepsilon$ in (8.4.21) and apply the Poincaré-Wirtinger inequality (Exercise 7.3.6) to show that $\|u_\varepsilon\|_{H^1(\Omega)}$ is uniformly bounded with respect to $\varepsilon > 0$. Thus, there is a subsequence, still denoted by $\{u_\varepsilon\}$, and some $u \in H^1(\Omega)$ such that $u_\varepsilon \rightharpoonup u$ in $H^1(\Omega)$ as $\varepsilon \to 0$.
(d) Recall the definition of the space V in (8.4.12). Show that the weak limit u from (c) satisfies (8.4.13). Since (8.4.13) admits a unique solution, the limit u does not depend on the subsequence chosen in (c). Hence, the entire family $\{u_\varepsilon\}$ converges weakly to u in $H^1(\Omega)$ as $\varepsilon \to 0$.
(e) Show that $\||\nabla u_\varepsilon|\|_{L^2(\Omega)} \to \||\nabla u|\|_{L^2(\Omega)}$ as $\varepsilon \to 0$, and conclude the strong convergence $u_\varepsilon \to u$ in $H^1(\Omega)$ as $\varepsilon \to 0$.

Exercise 8.4.8 Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz domain with boundary $\Gamma = \Gamma_D \cup \Gamma_N$, Γ_D relatively closed, Γ_N relatively open, $\Gamma_D \cap \Gamma_N = \emptyset$, and $\text{meas}(\Gamma_D) > 0$. Let $A \in L^\infty(\Omega)^{d\times d}$ be symmetric and uniformly positive definite, i.e.,

$$\begin{aligned} A(\boldsymbol{x})^T &= A(\boldsymbol{x}) \quad \text{a.e. } \boldsymbol{x} \in \Omega, \\ \boldsymbol{\xi}^T A(\boldsymbol{x})\, \boldsymbol{\xi} &\geq c_0 |\boldsymbol{\xi}|^2 \quad \text{a.e. } \boldsymbol{x} \in \Omega,\ \forall\, \boldsymbol{\xi} \in \mathbb{R}^d \end{aligned}$$

with some constant $c_0 > 0$, and let $c \in L^\infty(\Omega)$ be such that $c(\boldsymbol{x}) \geq 0$ a.e. $\boldsymbol{x} \in \Omega$. Given $f \in L^2(\Omega)$ and $g \in L^2(\Gamma_N)$, consider the boundary value problem

$$\begin{aligned} -\text{div}\,(A\nabla u) + c\, u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_D, \\ (A\nabla u) \cdot \boldsymbol{\nu} &= g \quad \text{on } \Gamma_N. \end{aligned}$$

(a) Derive the weak formulation and show the existence of a unique weak solution $u \in H^1_{\Gamma_D}(\Omega)$.
(b) Define the energy functional

$$E(v) = \int_\Omega \left[\frac{1}{2} \left(|A^{1/2} \nabla v|^2 + c\,|v|^2 \right) - f\,v \right] dx - \int_{\Gamma_N} g\,v\,ds, \quad v \in H^1_{\Gamma_D}(\Omega).$$

Compute the derivative $E'(v)$.
(c) Show that $u \in H^1_{\Gamma_D}(\Omega)$ is the solution of the weak formulation if and only if $E'(u) = 0$, and if and only if u minimizes $E(\cdot)$ over the space $H^1_{\Gamma_D}(\Omega)$.

Exercise 8.4.9 The biharmonic equation

$$\Delta^2 u = f \quad \text{in } \Omega$$

arises in fluid mechanics as well as thin elastic plate problems. Let us consider the biharmonic equation together with the homogeneous Dirichlet boundary conditions

$$u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma.$$

Note that the differential equation is of fourth-order, so boundary conditions involving the unknown function and first-order derivatives are treated as Dirichlet (or essential) boundary conditions, whereas Neumann (or natural) boundary conditions refer to those involving second and third order derivatives of the unknown function. Give a weak formulation of the homogeneous Dirichlet boundary value problem for the biharmonic equation and explore its unique solvability.

8.5 A boundary value problem of linearized elasticity

We study a boundary value problem of linearized elasticity in this section. Consider the deformation of an elastic body occupying a domain $\Omega \subset \mathbb{R}^d$; $d \leq 3$ in applications. The boundary Γ of the domain is assumed Lipschitz continuous so that the unit outward normal $\boldsymbol{\nu}$ exists almost everywhere on Γ. We divide the boundary Γ into two complementary parts Γ_D and Γ_N, where Γ_D is relatively closed, Γ_N is relatively open, $\Gamma_D \cap \Gamma_N = \emptyset$ and $\text{meas}(\Gamma_D) > 0$. The body is subject to the action of a body force of the density $\boldsymbol{f}$ and the surface traction of density $\boldsymbol{g}$ on Γ_N. We assume the body is fixed along Γ_D. As a result of the applications of the external forces, the body experiences some deformation and reaches an equilibrium state. A material point $\boldsymbol{x} \in \Omega$ in the undeformed body will be moved to the location $\boldsymbol{x} + \boldsymbol{u}$ after the deformation. The quantity $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x})$ is the *displacement* of the point $\boldsymbol{x}$. The displacement $\boldsymbol{u} : \Omega \subset \mathbb{R}^d \to \mathbb{R}^d$ is a vector-valued function, $d = 2$ for planar problems and $d = 3$ for three dimensional problems.

Besides the displacement, another important mechanical quantity is the *stress tensor*. Consider a surface within the body. Interaction between the

material on the two sides of the surface can be described in terms of forces exerted across the surface. A stress is the force per unit area exerted by the part of the body on one side of the surface on the part of the body on the other side. The state of the stress at any point in the body can be described by the stress tensor $\boldsymbol{\sigma}$. The stress tensor take on values in $\mathbb{S}^d$, the space of second order symmetric tensors on $\mathbb{R}^d$. We may simply view $\mathbb{S}^d$ as the space of symmetric matrices of order d. The canonical inner product and corresponding norm on $\mathbb{S}^d$ are

$$\boldsymbol{\sigma} : \boldsymbol{\tau} = \sigma_{ij}\tau_{ij}, \quad |\boldsymbol{\sigma}| = (\boldsymbol{\sigma} : \boldsymbol{\sigma})^{1/2} \quad \forall\, \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathbb{S}^d.$$

In discussing mechanical problems, we adopt the summation convention over a repeated index, i.e., if an index appears twice in an expression, a summation is implied with respect to that index. Thus, $\sigma_{ij}\tau_{ij}$ stands for $\sum_{i,j=1}^{d} \sigma_{ij}\tau_{ij}$.

Mathematical relations in a mechanical problem can be split into two kinds, material-independent ones, and material-dependent ones called *constitutive laws*. To describe these relations, we need the notion of a strain tensor. Strain is used to describe changes in size and shape of the body caused by the action of stress. The strain tensor $\boldsymbol{\varepsilon}$ quantifies the strain of the body undergoing the deformation. Like the stress tensor, the strain tensor takes on values in $\mathbb{S}^d$. A reader with little background on elasticity may simply view $\boldsymbol{u}$ as a d-dimensional vector-valued function, and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$ as $d \times d$ symmetric matrix-valued functions.

The material-independent relations include the strain-displacement relation, the equation of equilibrium, and boundary conditions. The equation of equilibrium takes the form

$$-\operatorname{div} \boldsymbol{\sigma} = \boldsymbol{f} \quad \text{in } \Omega. \tag{8.5.1}$$

Here $\operatorname{div} \boldsymbol{\sigma}$ is a d-dimensional vector-valued function with the components (note the summation over the index j)

$$(\operatorname{div} \boldsymbol{\sigma})_i = \sigma_{ij,j}, \quad 1 \le i \le d.$$

We assume the deformation is small (i.e., both the displacement and its gradient are small in size), and use the linearized strain tensor

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\left[\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^T\right] \quad \text{in } \Omega. \tag{8.5.2}$$

The specified boundary conditions take the form

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_D, \tag{8.5.3}$$

$$\boldsymbol{\sigma}\boldsymbol{\nu} = \boldsymbol{g} \quad \text{on } \Gamma_N. \tag{8.5.4}$$

Here $\boldsymbol{\sigma\nu}$ represents the action of the stress tensor $\boldsymbol{\sigma}$ along the direction of the unit outward normal $\boldsymbol{\nu}$. It can be viewed as a matrix-vector multiplication, and the result is a d-dimensional vector-valued function whose i^{th} component is $\sigma_{ij}\nu_j$.

The above relations are supplemented by a constitutive relation, which describes the mechanical response of the material to the external forces. The simplest constitutive relation is provided by that of linearized elasticity,

$$\boldsymbol{\sigma} = \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u}). \tag{8.5.5}$$

The elasticity tensor $\mathcal{C}$ is of fourth order, and can be viewed as a linear mapping from the space $\mathbb{S}^d$ to itself. With respect to the Cartesian coordinate system, the tensor $\mathcal{C}$ has the components C_{ijkl}, $1 \le i, j, k, l \le d$. The expression $\mathcal{C}\boldsymbol{\varepsilon}$ stands for a second-order tensor whose $(i, j)^{th}$ component is $C_{ijkl}\varepsilon_{kl}$. In component form, the constitutive relation (8.5.5) is rewritten as

$$\sigma_{ij} = C_{ijkl}\varepsilon_{kl}(\boldsymbol{u}), \quad 1 \le i, j \le d.$$

We assume the elasticity tensor $\mathcal{C}$ is bounded,

$$C_{ijkl} \in L^\infty(\Omega), \tag{8.5.6}$$

symmetric,

$$C_{ijkl} = C_{jikl} = C_{klij}, \tag{8.5.7}$$

and pointwise stable,

$$\boldsymbol{\varepsilon} : \mathcal{C}\boldsymbol{\varepsilon} \ge \alpha\, |\boldsymbol{\varepsilon}|^2 \quad \forall\, \boldsymbol{\varepsilon} \in \mathbb{S}^d \tag{8.5.8}$$

with a constant $\alpha > 0$.

If the components C_{ijkl} of the elasticity tensor do not depend on $\boldsymbol{x} \in \Omega$, the material is said to be *homogeneous*. Otherwise it is *nonhomogeneous*. For a fixed $\boldsymbol{x} \in \Omega$, if the tensor $\mathcal{C}(\boldsymbol{x})$ is invariant with respect to rotations of the coordinate system, the material is said to be *isotropic* at the point $\boldsymbol{x}$. Otherwise, the material is *anisotropic* at the point $\boldsymbol{x} \in \Omega$. For example, wood is anisotropic at any point since it possesses different properties along and across its fibers (such material is called *orthotropic*).

In the special case of an isotropic, homogeneous linearly elastic material, we have

$$C_{ijkl} = \lambda\, \delta_{ij}\delta_{kl} + \mu\, (\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \tag{8.5.9}$$

where $\lambda, \mu > 0$ are called Lamé moduli; then the constitutive relation is simplified to

$$\boldsymbol{\sigma} = \lambda\, (\operatorname{tr} \boldsymbol{\varepsilon})\, \boldsymbol{I} + 2\, \mu\, \boldsymbol{\varepsilon}. \tag{8.5.10}$$

Here we use $\boldsymbol{I}$ to denote the unit tensor of the second order (think of it as the identity matrix of order d), and $\operatorname{tr} \boldsymbol{\varepsilon}$ is the trace of the tensor (matrix) $\boldsymbol{\varepsilon}$. In component form,

$$\sigma_{ij} = \lambda\, (\varepsilon_{kk})\, \delta_{ij} + 2\, \mu\, \varepsilon_{ij}. \tag{8.5.11}$$

Inverting (8.5.11) we get

$$\varepsilon_{ij} = \frac{1+\nu}{E}\,\sigma_{ij} - \frac{\nu}{E}\,(\sigma_{kk})\,\delta_{ij} \tag{8.5.12}$$

where the constants E and ν are defined by the relations

$$E = \frac{\mu(3\lambda+2\mu)}{\lambda+\mu}, \quad \nu = \frac{\lambda}{2(\lambda+\mu)}\,.$$

Young's modulus E measures the stiffness in the tension (axial) direction, and *Poisson's ratio* ν measures the lateral contraction. Another quantity used often is the *bulk modulus* K, defined by

$$K = \frac{E}{3\,(1-2\,\nu)} = \frac{3\lambda+2\mu}{3}.$$

These quantities are used in the engineering literature more often than the Lamé coefficients λ and μ. Some experimental tests (see, e.g. [177]) lead to restrictions $\lambda > 0$, $\mu > 0$ for the Lamé coefficients and it follows that $K > 0$, $E > 0$, $0 < \nu < 1/2$.

We now introduce a useful integration by parts formula. For any symmetric tensor $\boldsymbol{\sigma}$ and any vector field $\boldsymbol{v}$, both being continuously differentiable over $\overline{\Omega}$,

$$\int_\Omega \operatorname{div}\boldsymbol{\sigma}\cdot\boldsymbol{v}\,dx = \int_\Gamma \boldsymbol{\sigma}\boldsymbol{\nu}\cdot\boldsymbol{v}\,ds - \int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{v})\,dx. \tag{8.5.13}$$

The classical formulation of the boundary value problem for the linearized elasticity consists of the equations (8.5.1)–(8.5.5). We now derive the corresponding weak formulation with regard to the unknown variable $\boldsymbol{u}$. Suppose the linearized elasticity problem has a smooth solution $\boldsymbol{u}$. By (8.5.2) and (8.5.5), $\boldsymbol{\sigma}$ is also smooth. Combining the equations (8.5.1) and (8.5.2), we see that the differential equation is of second order for $\boldsymbol{u}$. We multiply the equation (8.5.1) by an arbitrary smooth test function $\boldsymbol{v}$ that vanishes on Γ_D, and integrate over Ω,

$$-\int_\Omega \operatorname{div}\boldsymbol{\sigma}\cdot\boldsymbol{v}\,dx = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx.$$

Apply the formula (8.5.13) to the left hand side,

$$-\int_\Gamma (\boldsymbol{\sigma}\boldsymbol{\nu})\cdot\boldsymbol{v}\,ds + \int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{v})\,dx = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx.$$

Upon the use of the boundary conditions, the boundary integral term can be written as

$$-\int_\Gamma (\boldsymbol{\sigma}\boldsymbol{\nu})\cdot\boldsymbol{v}\,ds = -\int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{v}\,ds.$$

Therefore,

$$\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx + \int_{\Gamma_N} \boldsymbol{g} \cdot \boldsymbol{v}\, ds.$$

Recalling the constitutive law (8.5.5), we have thus derived the following relation for any smooth function $\boldsymbol{v}$ that vanishes on Γ_D:

$$\int_\Omega [\mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u})] : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx + \int_{\Gamma_N} \boldsymbol{g} \cdot \boldsymbol{v}\, ds.$$

Assume

$$\boldsymbol{f} \in [L^2(\Omega)]^d, \quad \boldsymbol{g} \in [L^2(\Gamma_N)]^d \tag{8.5.14}$$

and introduce the function space

$$V = \left\{ \boldsymbol{v} \in [H^1(\Omega)]^d \mid \boldsymbol{v} = \boldsymbol{0} \text{ a.e. on } \Gamma_D \right\}.$$

Then the weak formulation for the displacement variable is

$$\boldsymbol{u} \in V, \quad \int_\Omega [\mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u})] : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx + \int_{\Gamma_N} \boldsymbol{g} \cdot \boldsymbol{v}\, ds \quad \forall\, \boldsymbol{v} \in V. \tag{8.5.15}$$

We can apply the Lax-Milgram Lemma to conclude the existence and uniqueness of the problem (8.5.15).

Theorem 8.5.1 *Assume* (8.5.6)–(8.5.8), (8.5.14) *and* $\operatorname{meas}(\Gamma_D) > 0$. *Then there is a unique solution to the problem* (8.5.15), *and the problem* (8.5.15) *is equivalent to the minimization problem*

$$\boldsymbol{u} \in V, \quad E(\boldsymbol{u}) = \inf\{E(\boldsymbol{v}) \mid \boldsymbol{v} \in V\}, \tag{8.5.16}$$

where the energy functional is defined by

$$E(\boldsymbol{v}) = \frac{1}{2} \int_\Omega [\mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{v})] : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx - \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx - \int_{\Gamma_N} \boldsymbol{g} \cdot \boldsymbol{v}\, ds. \tag{8.5.17}$$

A proof of this result is left as Exercise 8.5.3. In verifying the V-ellipticity of the bilinear form

$$a(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \left[\mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u})\right] : \boldsymbol{\varepsilon}(\boldsymbol{v})\, dx,$$

we need to apply Korn's inequality: There exists a constant $c > 0$ depending only on Ω such that

$$\|\boldsymbol{v}\|^2_{[H^1(\Omega)]^d} \le c \int_\Omega |\boldsymbol{\varepsilon}(\boldsymbol{v})|^2 dx \quad \forall\, \boldsymbol{v} \in V.$$

Exercise 8.5.1 Prove the integration by part formula (8.5.13).

Exercise 8.5.2 In the case of an isotropic, homogeneous linearly elastic material (8.5.10), show that the classical formulation of the equilibrium equation written in terms of the displacement is

$$-\mu\,\Delta\boldsymbol{u} - (\lambda+\mu)\,\nabla\mathrm{div}\,\boldsymbol{u} = \boldsymbol{f} \quad \text{in } \Omega.$$

Give a derivation of the weak formulation of the problem: Find $\boldsymbol{u} \in V$ such that

$$a(\boldsymbol{u},\boldsymbol{v}) = \ell(\boldsymbol{v}) \quad \forall\,\boldsymbol{v} \in V,$$

where

$$V = \left\{\boldsymbol{v} \in [H^1(\Omega)]^d \mid \boldsymbol{v} = \boldsymbol{0} \text{ on } \Gamma_D\right\},$$
$$a(\boldsymbol{u},\boldsymbol{v}) = \int_\Omega [\lambda\,\mathrm{div}\,\boldsymbol{u}\,\mathrm{div}\,\boldsymbol{v} + 2\,\mu\,\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})]\,dx,$$
$$\ell(\boldsymbol{v}) = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx + \int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{v}\,ds.$$

Prove that the weak formulation has a unique solution.

Exercise 8.5.3 Apply the Lax-Milgram Lemma to prove Theorem 8.5.1.

Exercise 8.5.4 This exercise follows [92]. In solving the Navier-Stokes equations for incompressible viscous flows by some operator splitting schemes, such as Peaceman-Rachford scheme, at each time step, one obtains a subproblem of the following type: Find $\boldsymbol{u} \in H^1(\Omega)^d$, $\boldsymbol{u} = \boldsymbol{g}_0$ on Γ_0 such that

$$a(\boldsymbol{u},\boldsymbol{v}) = \ell(\boldsymbol{v}) \quad \forall\,\boldsymbol{v} \in H^1_{\Gamma_0}(\Omega)^d. \tag{8.5.18}$$

Here, Ω is a Lipschitz domain in $\mathbb{R}^d$, $\partial\Omega = \Gamma_0 \cup \Gamma_1$ with Γ_0 relatively closed, Γ_1 relatively open, and $\Gamma_0 \cap \Gamma_1 = \emptyset$. The function $\boldsymbol{g}_0 \in H^1(\Omega)^d$ is given. The bilinear form is

$$a(\boldsymbol{u},\boldsymbol{v}) = \alpha\int_\Omega \boldsymbol{u}\cdot\boldsymbol{v}\,dx + \mu\int_\Omega \nabla\boldsymbol{u} : \nabla\boldsymbol{v}\,dx + \int_\Omega (\boldsymbol{V}\cdot\nabla)\boldsymbol{u}\cdot\boldsymbol{v}\,dx,$$

where α and μ are given positive constants, $\boldsymbol{V} \in H^1(\Omega)^d \cap L^\infty(\Omega)^d$ ($\boldsymbol{V} \in H^1(\Omega)^d$ for $d \le 4$) is a given function satisfying

$$\mathrm{div}\,\boldsymbol{V} = 0 \text{ in } \Omega, \quad \boldsymbol{V}\cdot\boldsymbol{\nu} \ge 0 \text{ on } \Gamma_1.$$

The linear form is

$$\ell(\boldsymbol{v}) = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx + \int_{\Gamma_1} \boldsymbol{g}_1\cdot\boldsymbol{v}\,ds,$$

where $\boldsymbol{f} \in L^2(\Omega)^d$ and $\boldsymbol{g}_1 \in L^2(\Gamma_1)^d$ are given functions. Denote the bilinear form

$$b(\boldsymbol{v},\boldsymbol{w}) = \int_\Omega (\boldsymbol{V}\cdot\nabla)\boldsymbol{v}\cdot\boldsymbol{w}\,dx, \quad \boldsymbol{v},\boldsymbol{w} \in H^1(\Omega)^d.$$

Show the following:
(a) $b(\cdot,\cdot)$ is continuous over $H^1(\Omega)^d \times H^1(\Omega)^d$.
(b) $b(\boldsymbol{v},\boldsymbol{v}) \ge 0$ for any $\boldsymbol{v} \in H^1_{\Gamma_0}(\Omega)^d$.

Hint: The space $C^\infty_{\Gamma_0}(\overline{\Omega}) = \{v \in C^\infty(\overline{\Omega}) \mid v = 0 \text{ in a neighborhood of } \Gamma_0\}$ is dense in $H^1_{\Gamma_0}(\Omega)$.
(c) The problem (8.5.18) has a unique solution.
(d) Derive the classical formulation of the problem (8.5.18).

Note: The problem (8.5.18) is a linear advection-diffusion problem. The assumption "$\boldsymbol{V} \cdot \boldsymbol{\nu} \geq 0$ on Γ_1" reflects the fact that the Neumann boundary condition is specified on a downstream part Γ_1 of the flow region boundary.

8.6 Mixed and dual formulations

This section is intended as a brief introduction to two more weak formulations for boundary value problems, namely the mixed formulation and the dual formulation. We use the model problem

$$-\Delta u = f \quad \text{in } \Omega, \tag{8.6.1}$$

$$u = 0 \quad \text{on } \Gamma \tag{8.6.2}$$

for a description of the new weak formulations. Assume $f \in L^2(\Omega)$. The weak formulation discussed in previous sections is

$$u \in H^1_0(\Omega), \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx \quad \forall \, v \in H^1_0(\Omega). \tag{8.6.3}$$

This weak formulation is called the *primal formulation* since the unknown variable is u. Here the bilinear form is symmetric, so the weak formulation (8.6.3) is equivalent to the minimization problem

$$u \in H^1_0(\Omega), \quad E(u) = \inf_{v \in H^1_0(\Omega)} E(v) \tag{8.6.4}$$

with

$$E(v) = \frac{1}{2} \int_\Omega |\nabla v|^2 dx - \int_\Omega f \, v \, dx. \tag{8.6.5}$$

In the context of a heat conduction problem, u is the temperature and ∇u has a physical meaning related to the heat flux. In many situations, the heat flux is a more important quantity than the temperature variable. It is then desirable to develop equivalent weak formulations of the boundary value problem (8.6.1)–(8.6.2) that involves $\boldsymbol{p} = \nabla u$ as an unknown. For this purpose, let $\boldsymbol{q} = \nabla v$. Then noticing that

$$\frac{1}{2} \int_\Omega |\nabla v|^2 dx = \sup_{\boldsymbol{q} \in L^2(\Omega)^d} \int_\Omega \Big(\boldsymbol{q} \cdot \nabla v - \frac{1}{2} |\boldsymbol{q}|^2 \Big) dx,$$

we can replace the minimization problem (8.6.4) by

$$\inf_{v \in H^1_0(\Omega)} \sup_{\boldsymbol{q} \in L^2(\Omega)^d} L(v, \boldsymbol{q}), \tag{8.6.6}$$

where

$$L(v, \boldsymbol{q}) = \int_\Omega \left(\boldsymbol{q} \cdot \nabla v - \frac{1}{2} |\boldsymbol{q}|^2 - f\, v \right) dx \tag{8.6.7}$$

is a Lagrange function. We now consider a new problem

$$\sup_{\boldsymbol{q} \in L^2(\Omega)^d} \inf_{v \in H_0^1(\Omega)} L(v, \boldsymbol{q}), \tag{8.6.8}$$

obtained by interchanging inf and sup in (8.6.6). We have

$$\inf_{v \in H_0^1(\Omega)} L(v, \boldsymbol{q}) = \begin{cases} -\dfrac{1}{2} \displaystyle\int_\Omega |\boldsymbol{q}|^2 dx & \text{if } \boldsymbol{q} \in Q_f, \\ -\infty & \text{if } \boldsymbol{q} \notin Q_f, \end{cases}$$

where

$$Q_f = \left\{ \boldsymbol{q} \in L^2(\Omega)^d \;\middle|\; \int_\Omega \boldsymbol{q} \cdot \nabla v \, dx = \int_\Omega f\, v \, dx \;\forall\, v \in H_0^1(\Omega) \right\}.$$

The constraint in Q_f is the weak form of the relation $-\operatorname{div} \boldsymbol{q} = f$. Thus the problem (8.6.8) is equivalent to

$$\sup_{\boldsymbol{q} \in Q_f} \left(-\frac{1}{2} \int_\Omega |\boldsymbol{q}|^2 dx \right) \tag{8.6.9}$$

or

$$\inf_{\boldsymbol{q} \in Q_f} \frac{1}{2} \int_\Omega |\boldsymbol{q}|^2 dx.$$

Problem (8.6.9) is called a dual formulation of the model problem (8.6.1)–(8.6.2), whereas problem (8.6.4) is called the primal formulation of the model problem.

Relation between a primal formulation and its associated dual formulation can be studied within the theory of saddle point problems. A general presentation of the theory is found in [76, Chap. VI]. Here we quote the main result.

Definition 8.6.1 *Let A and B be two sets and $L : A \times B \to \mathbb{R}$. Then a pair $(u, p) \in A \times B$ is said to be a saddle point of L if*

$$L(u, q) \le L(u, p) \le L(v, p) \quad \forall\, (v, q) \in A \times B. \tag{8.6.10}$$

This definition generalizes the concept of a saddle point of a real-valued function of two real variables. Recall that $(1, 2)$ is a saddle point of the function

$$y = (x_1 - 1)^2 - (x_2 - 2)^2, \quad x_1, x_2 \in \mathbb{R}.$$

The problem (8.6.10) of finding a saddle point of the functional L is called a *saddle point problem*. The next result follows immediately from Definition 8.6.1.

Proposition 8.6.2 *A pair $(u,p) \in A \times B$ is a saddle point of $L : A \times B \to \mathbb{R}$ if and only if*

$$\max_{q\in B} \inf_{v\in A} L(v,q) = \min_{v\in A} \sup_{q\in B} L(v,q), \tag{8.6.11}$$

and the common value equals $L(u,p)$.

Note that with a saddle point (u,p), the maximum in (8.6.11) is attained at $q = p$, and the minimum is attained at $v = u$.

Regarding existence and uniqueness of a saddle point, we have the following general result.

Theorem 8.6.3 *Assume*
(a) *V and Q are reflexive Banach spaces;*
(b) *$A \subset V$ is non-empty, closed and convex;*
(c) *$B \subset Q$ is non-empty, closed and convex;*
(d) *$\forall\, q \in B$, $v \mapsto L(v,q)$ is convex and l.s.c. on A;*
(e) *$\forall\, v \in A$, $q \mapsto L(v,q)$ is concave and u.s.c. on B;*
(f) *$A \subset V$ is bounded, or $\exists\, q_0 \in B$ such that*

$$L(v,q_0) \to \infty \quad \text{as } \|v\| \to \infty,\ v \in A;$$

(g) *$B \subset Q$ is bounded, or $\exists\, v_0 \in A$ such that*

$$L(v_0,q) \to -\infty \quad \text{as } \|q\| \to \infty,\ q \in B.$$

Then L has a saddle point $(u,p) \in A \times B$. Moreover, if the convexity in (d) *is strengthened to strict convexity, then the first component u is unique. Similarly, if the concavity in* (e) *is strengthened to strict concavity, then the second component p is unique.*

The same conclusions hold when (f) *is replaced by*
(f)′ *$A \subset V$ is bounded, or*

$$\lim_{\substack{\|v\|\to\infty \\ v\in A}} \sup_{q\in B} L(v,q) = \infty;$$

or when (g) *is replaced by*
(g)′ *$B \subset Q$ is bounded, or*

$$\lim_{\substack{\|q\|\to\infty \\ q\in B}} \inf_{v\in A} L(v,q) = -\infty.$$

In the statement of Theorem 8.6.3, we use the notions of concavity and upper semi-continuity. A function f is said to be concave if $(-f)$ is convex, and f is u.s.c. (upper semi-continuous) if $(-f)$ is l.s.c.

Corresponding to the saddle point problem (8.6.10), define two functionals:

$$E(v) = \sup_{q\in B} L(v,q), \quad v \in A,$$
$$E^c(q) = \inf_{v\in A} L(v,q), \quad q \in B.$$

We call

$$\inf_{v\in A} E(v) \tag{8.6.12}$$

the primal problem, and

$$\sup_{q\in B} E^c(q) \tag{8.6.13}$$

the dual problem. Combining Theorem 8.6.3 and Proposition 8.6.2 we see that under the assumptions listed in Theorem 8.6.3, the primal problem (8.6.12) has a solution $u \in A$, the dual problem (8.6.13) has a solution $p \in B$, and

$$\inf_{v\in A} E(v) = \sup_{q\in B} E^c(q) = L(u,p). \tag{8.6.14}$$

Back to the model problem (8.6.1)–(8.6.2), we have the primal formulation (8.6.4), the dual formulation (8.6.9), the Lagrange functional $L(v,\boldsymbol{q})$ given by (8.6.7), $A = V = H_0^1(\Omega)$, and $B = Q = L^2(\Omega)^d$. It is easy to see that the assumptions (a)–(e) and (g) in Theorem 8.6.3 are satisfied. We can verify (f)′ as follows:

$$\sup_{\boldsymbol{q}\in L^2(\Omega)^d} L(v,\boldsymbol{q}) \geq L(v,\nabla v) = \int_\Omega \Big(\frac{1}{2}\,|\nabla v|^2 - f\,v\Big)\,dx$$

which tends to ∞ as $\|v\|_{H_0^1(\Omega)} \to \infty$. Note that for any $v \in V$, the mapping $\boldsymbol{q} \mapsto L(v,\boldsymbol{q})$ is strictly concave. Also, the primal formulation (8.6.4) has a unique solution. We thus conclude that the Lagrange functional $L(v,\boldsymbol{q})$ of (8.6.7) has a unique saddle point $(u,\boldsymbol{p}) \in H_0^1(\Omega)\times L^2(\Omega)^d$:

$$L(u,\boldsymbol{q}) \leq L(u,\boldsymbol{p}) \leq L(v,\boldsymbol{p}) \quad \forall\,\boldsymbol{q}\in L^2(\Omega)^d,\ \forall\, v\in H_0^1(\Omega), \tag{8.6.15}$$

and

$$E(u) = \inf_{v\in H_0^1(\Omega)} E(v) = \sup_{\boldsymbol{q}\in Q_f} \left(-\frac{1}{2}\int_\Omega |\boldsymbol{q}|^2 dx\right) = -\frac{1}{2}\int_\Omega |\boldsymbol{p}|^2 dx.$$

It is left as an exercise to show that the inequalities (8.6.15) are equivalent to $(u,\boldsymbol{p}) \in H_0^1(\Omega)\times L^2(\Omega)^d$ satisfying

$$\int_\Omega \boldsymbol{p}\cdot\boldsymbol{q}\,dx - \int_\Omega \boldsymbol{q}\cdot\nabla u\,dx = 0 \qquad \forall\,\boldsymbol{q}\in L^2(\Omega)^d, \tag{8.6.16}$$

$$-\int_\Omega \boldsymbol{p}\cdot\nabla v\,dx = -\int_\Omega f\,v\,dx \qquad \forall\, v\in H_0^1(\Omega). \tag{8.6.17}$$

Upon an integration by parts, another weak formulation is obtained: Find $(u,\boldsymbol{p}) \in L^2(\Omega)\times H(\mathrm{div};\Omega)$ such that

$$\int_\Omega \boldsymbol{p}\cdot\boldsymbol{q}\,dx + \int_\Omega (\mathrm{div}\,\boldsymbol{q})\,u\,dx = 0 \qquad \forall\,\boldsymbol{q}\in H(\mathrm{div};\Omega), \tag{8.6.18}$$

$$\int_\Omega (\mathrm{div}\,\boldsymbol{p})\,v\,dx = -\int_\Omega f\,v\,dx \qquad \forall\, v\in L^2(\Omega). \tag{8.6.19}$$

Here

$$H(\mathrm{div};\Omega) = \left\{ \boldsymbol{q} \in L^2(\Omega)^d \mid \mathrm{div}\, \boldsymbol{q} \in L^2(\Omega) \right\}; \tag{8.6.20}$$

it is a Hilbert space with the inner product

$$(\boldsymbol{p}, \boldsymbol{q})_{H(\mathrm{div};\Omega)} = (\boldsymbol{p}, \boldsymbol{q})_{L^2(\Omega)^d} + (\mathrm{div}\, \boldsymbol{p}, \mathrm{div}\, \boldsymbol{q})_{L^2(\Omega)}.$$

Formulations (8.6.16)–(8.6.17) and (8.6.18)–(8.6.19) are examples of *mixed formulations* and they fall in the following abstract framework:

Let V and Q be two Hilbert spaces. Assume $a(u, v)$ is a continuous bilinear form on $V \times V$, and $b(v, q)$ is a continuous bilinear form on $V \times Q$. Given $f \in V'$ and $g \in Q'$, find $u \in V$ and $p \in Q$ such that

$$a(u, v) + b(v, p) = \langle f, v \rangle_{V' \times V} \quad \forall\, v \in V, \tag{8.6.21}$$

$$b(u, q) = \langle g, q \rangle_{Q' \times Q} \quad \forall\, q \in Q. \tag{8.6.22}$$

In addition to the need of bringing in quantities of physical importance into play in a weak formulation (e.g. $\boldsymbol{p} = \nabla u$ for the model problem) in the hope of achieving more accurate numerical approximations for them, we frequently arrive at mixed formulation of the type (8.6.21)–(8.6.22) in dealing with constraints such as the incompressibility in fluids or in certain solids (see Exercise 8.6.4). Finite element approximations based on mixed formulations are called *mixed finite element methods*, and they have been extensively analyzed and applied in solving mechanical problems. In Chapter 10, we only discuss the finite element method based on the primal weak formulation. The interested reader can consult [43] for a detailed discussion of the mixed formulation (8.6.21)–(8.6.22) and mixed finite element methods.

Exercise 8.6.1 Show the equivalence between (8.6.15) and (8.6.16)–(8.6.17).

Exercise 8.6.2 Show that (8.6.16) and (8.6.17) are the weak formulation of the equations

$$\boldsymbol{p} = \nabla u, \quad -\mathrm{div}\, \boldsymbol{p} = f.$$

Exercise 8.6.3 Show that the weak formulation (8.6.18)–(8.6.19) is equivalent to the saddle point problem for

$$L(v, \boldsymbol{q}) = \int_\Omega \left[-(f + \mathrm{div}\, \boldsymbol{q})\, v - \frac{1}{2} |\boldsymbol{q}|^2 \right] dx$$

on $L^2(\Omega) \times H(\mathrm{div};\Omega)$.

Exercise 8.6.4 As an example of a mixed formulation for the treatment of a constraint, we consider the following boundary value problem of the Stokes equations for an incompressible Newtonian fluid with viscosity $\mu > 0$: Given a force

density $\boldsymbol{f}$, find a velocity $\boldsymbol{u}$ and a pressure p such that

$$\begin{aligned} -\mu\,\Delta\boldsymbol{u} + \nabla p &= \boldsymbol{f} \quad \text{in } \Omega, \\ \operatorname{div}\boldsymbol{u} &= 0 \quad \text{in } \Omega, \\ \boldsymbol{u} &= \boldsymbol{0} \quad \text{on } \Gamma. \end{aligned}$$

To uniquely determine p, we impose the condition

$$\int_\Omega p\,dx = 0.$$

Assume $\boldsymbol{f} \in L^2(\Omega)^d$. Let

$$V = H_0^1(\Omega)^d, \quad Q = \left\{ q \in L^2(\Omega) \;\middle|\; \int_\Omega q\,dx = 0 \right\}.$$

Show that a mixed weak formulation of the boundary value problem is: Find $\boldsymbol{u} = (u_1, \ldots, u_d)^T \in V$ and $p \in Q$ such that

$$\begin{aligned} \mu \sum_{i=1}^d \int_\Omega \nabla u_i \cdot \nabla v_i\,dx - \int_\Omega p \operatorname{div}\boldsymbol{v}\,dx &= \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx \quad \forall\,\boldsymbol{v} \in V, \\ \int_\Omega q \operatorname{div}\boldsymbol{u}\,dx &= 0 \quad \forall\, q \in Q. \end{aligned}$$

Introduce the Lagrange functional

$$L(\boldsymbol{v}, q) = \int_\Omega \left[\mu\,\boldsymbol{\varepsilon}(\boldsymbol{v}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) - q \operatorname{div}\boldsymbol{v} - \boldsymbol{f}\cdot\boldsymbol{v}\right] dx,$$

where $\boldsymbol{\varepsilon}(\boldsymbol{v})$ is the linearized strain tensor defined in (8.5.2). Show that $(\boldsymbol{u}, p) \in V \times Q$ satisfies the weak formulation if and only if it is a saddle point of $L(\boldsymbol{v}, q)$ over $V \times Q$.

8.7 Generalized Lax-Milgram Lemma

The following result extends Lax-Milgram Lemma, and is due to Nečas [176].

Theorem 8.7.1 *Let U and V be real Hilbert spaces, $a : U \times V \to \mathbb{R}$ be a bilinear form, and $\ell \in V'$. Assume there are constants $M > 0$ and $\alpha > 0$ such that*

$$|a(u, v)| \le M\,\|u\|_U \|v\|_V \quad \forall\, u \in U,\ v \in V, \tag{8.7.1}$$

$$\sup_{0 \neq v \in V} \frac{a(u, v)}{\|v\|_V} \ge \alpha\,\|u\|_U \quad \forall\, u \in U, \tag{8.7.2}$$

$$\sup_{u \in U} a(u, v) > 0 \quad \forall\, v \in V,\ v \neq 0. \tag{8.7.3}$$

Then there exists a unique solution u of the problem

$$u \in U, \quad a(u,v) = \ell(v) \quad \forall\, v \in V. \tag{8.7.4}$$

Moreover,

$$\|u\|_U \le \frac{\|\ell\|_{V'}}{\alpha}. \tag{8.7.5}$$

Proof. The proof is similar to the second proof of Lax-Milgram Lemma, and we apply Theorem 8.2.1.

Again, let $A : U \to V$ be the linear continuous operator defined by the relation

$$a(u,v) = (Au, v)_V \quad \forall\, u \in U,\ v \in V.$$

Using the condition (8.7.1), we have

$$\|Au\|_V \le M \|u\|_U \quad \forall\, u \in U.$$

Then the problem (8.7.4) can be rewritten as

$$u \in U, \quad Au = \mathcal{J}\ell, \tag{8.7.6}$$

where $\mathcal{J} : V' \to V$ is the Riesz isometric operator.

From the condition (8.7.2) and the definition of A, it follows immediately that A is injective, i.e., $Au = 0$ for some $u \in U$ implies $u = 0$.

To show that the range $\mathcal{R}(A)$ is closed, let $\{u_n\} \subset U$ be a sequence such that $\{Au_n\}$ converges in V, the limit being denoted by $w \in V$. Using the condition (8.7.2), we have

$$\|u_m - u_n\|_U \le \frac{1}{\alpha} \sup_{0 \ne v \in V} \frac{(A(u_m - u_n), v)_V}{\|v\|_V} \le \frac{1}{\alpha} \|Au_m - Au_n\|_V.$$

Hence, $\{u_n\}$ is a Cauchy sequence in U, and hence have a limit $u \in U$. Moreover by the continuity condition (8.7.1), $Au_n \to Au = w$ in V. Thus, the range $\mathcal{R}(A)$ is closed.

Now if $v \in \mathcal{R}(A)^\perp$, then

$$(Au, v)_V = a(u,v) = 0 \quad \forall\, u \in U.$$

Applying the condition (8.7.3), we conclude $v = 0$. So $\mathcal{R}(A)^\perp = \{0\}$.

Therefore, the equation (8.7.6) and hence also the problem (8.7.4) has a unique solution.

The estimate (8.7.5) follows easily from another application of the condition (8.7.2). □

Exercise 8.7.1 Show that Theorem 8.7.1 is a generalization of the Lax-Milgram Lemma.

Exercise 8.7.2 As an application of Theorem 8.7.1, we consider the model boundary value problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma. \end{aligned}$$

The "standard" weak formulation of the problem is

$$u \in H_0^1(\Omega), \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \langle f, v \rangle_{H^{-1}(\Omega) \times H_0^1(\Omega)} \quad \forall\, v \in H_0^1(\Omega).$$

This formulation makes sense as long as $f \in H^{-1}(\Omega)$ (e.g. if $f \in L^2(\Omega)$). Performing an integration by part on the bilinear form, we are led to a new weak formulation:

$$u \in L^2(\Omega), \quad -\int_\Omega u \, \Delta v \, dx = \langle f, v \rangle_{(H^2(\Omega))' \times H^2(\Omega)} \quad \forall\, v \in H^2(\Omega) \cap H_0^1(\Omega). \tag{8.7.7}$$

This formulation makes sense even when $f \notin H^{-1}(\Omega)$ as long as $f \in [H^2(\Omega)]'$. One example is for $d \leq 3$, the point load

$$f(\boldsymbol{x}) = c_0 \delta(\boldsymbol{x} - \boldsymbol{x}_0)$$

for some $c_0 \in \mathbb{R}$ and $\boldsymbol{x}_0 \in \Omega$. In this case, we interpret $\langle f, v \rangle_{[H^2(\Omega)]' \times H^2(\Omega)}$ as $c_0 v(\boldsymbol{x}_0)$, that is well-defined since $H^2(\Omega)$ is embedded in $C(\overline{\Omega})$ when $d \leq 3$.

Assume $f \in (H^2(\Omega))'$ and $\Omega \subset \mathbb{R}^d$ is smooth or convex. Apply Theorem 8.7.1 to show that there is a unique "weaker" solution $u \in L^2(\Omega)$ to the problem (8.7.7). In verifying the condition (8.7.2), apply the estimate (7.3.11).

8.8 A nonlinear problem

A number of physical applications lead to partial differential equations of the type (see [246])

$$-\operatorname{div}\left[\alpha(|\nabla u|)\, \nabla u\right] = f.$$

In this section, we consider one such nonlinear equation. Specifically, we study the boundary value problem

$$-\operatorname{div}\left[(1 + |\nabla u|^2)^{p/2-1} \nabla u\right] = f \quad \text{in } \Omega, \tag{8.8.1}$$

$$u = 0 \quad \text{on } \Gamma, \tag{8.8.2}$$

where $p \in (1, \infty)$. We use p^* to denote the conjugate exponent defined through the relation

$$\frac{1}{p} + \frac{1}{p^*} = 1.$$

When $p = 2$, (8.8.1)–(8.8.2) reduces to a linear problem: the homogeneous Dirichlet boundary value problem for the Poisson equation, which was studied in Section 8.1.

Let us first formally derive a weak formulation for the problem (8.8.1)–(8.8.2). For this, we assume the problem has a solution, sufficiently smooth so that all the following calculations leading to the weak formulation are meaningful. Multiplying the equation (8.8.1) with an arbitrary test function $v \in C_0^\infty(\Omega)$ and integrating the relation over Ω, we have

$$-\int_\Omega \operatorname{div}\left[(1+|\nabla u|^2)^{p/2-1}\nabla u\right] v\,dx = \int_\Omega f\,v\,dx.$$

Then perform an integration by parts to obtain

$$\int_\Omega (1+|\nabla u|^2)^{p/2-1}\nabla u\cdot\nabla v\,dx = \int_\Omega f\,v\,dx. \tag{8.8.3}$$

Let us introduce the space

$$V = W_0^{1,p}(\Omega) \tag{8.8.4}$$

and define the norm

$$\|v\|_V = \Big(\int_\Omega |\nabla v|^p dx\Big)^{1/p}. \tag{8.8.5}$$

It can be verified that $\|\cdot\|_V$ defined in (8.8.5) is a norm over the space V, which is equivalent to the standard norm $\|\cdot\|_{W^{1,p}(\Omega)}$ (see Exercise 8.8.1). Since $p \in (1,\infty)$, the space V is a reflexive Banach space. The dual space of V is

$$V' = W^{-1,p^*}(\Omega)$$

and we assume throughout this section that

$$f \in V'.$$

Notice that the dual space V' is pretty large, e.g. $L^{p^*}(\Omega) \subset V'$.

It can be shown that the left side of (8.8.3) makes sense as long as $u, v \in V$ (see Exercise 8.8.2). Additionally, the right side of (8.8.3) is well defined for $f \in V'$ and $v \in V$ when we interpret the right side of (8.8.3) as the duality pair between V^* and V.

Now we are ready to introduce the weak formulation for the boundary value problem (8.8.1)–(8.8.2):

$$u \in V, \quad a(u;u,v) = \ell(v) \quad \forall\, v \in V. \tag{8.8.6}$$

Here

$$a(w;u,v) = \int_\Omega (1+|\nabla w|^2)^{p/2-1}\nabla u\cdot\nabla v\,dx, \tag{8.8.7}$$

$$\ell(v) = \int_\Omega f\,v\,dx. \tag{8.8.8}$$

Related to the weak formulation (8.8.6), we introduce a minimization problem

$$u \in V, \quad E(u) = \inf_{v \in V} E(v), \tag{8.8.9}$$

where the "energy functional" $E(\cdot)$ is

$$E(v) = \frac{1}{p} \int_\Omega (1 + |\nabla v|^2)^{p/2} dx - \int_\Omega f\, v\, dx. \tag{8.8.10}$$

We first explore some properties of the energy functional.

Lemma 8.8.1 *The energy functional $E(\cdot)$ is coercive, i.e.,*

$$E(v) \to \infty \quad \text{as } \|v\|_V \to \infty.$$

Proof. It is easy to see that

$$E(v) \geq \frac{1}{p} \|v\|_V^p - \|f\|_{V'} \|v\|_V.$$

Since $p > 1$, we have $E(v) \to \infty$ as $\|v\|_V \to \infty$. □

Lemma 8.8.2 *The energy functional $E(\cdot)$ is continuous.*

Proof. For $u, v \in V$, consider the difference

$$E(v) - E(u) = \frac{1}{p} \int_\Omega \left[(1 + |\nabla v|^2)^{p/2} - (1 + |\nabla u|^2)^{p/2} \right] dx - \int_\Omega f\, (v - u)\, dx.$$

Introduce the real variable function

$$g(t) = \frac{1}{p} \left[1 + |\nabla u + t\, \nabla (v - u)|^2 \right]^{p/2}, \quad 0 \leq t \leq 1.$$

We have

$$g(1) - g(0) = \int_0^1 g'(t)\, dt$$

with

$$g'(t) = \left[1 + |\nabla u + t\, \nabla (v - u)|^2 \right]^{p/2 - 1} \left[\nabla u + t\, \nabla (v - u) \right] \cdot \nabla (v - u).$$

So

$$\begin{aligned} |g(1) - g(0)| &\leq \int_0^1 \left[1 + |\nabla u + t\, \nabla (v - u)|^2 \right]^{(p-1)/2} |\nabla (v - u)|\, dt \\ &\leq \left[1 + (|\nabla u| + |\nabla v|)^2 \right]^{(p-1)/2} |\nabla (v - u)|. \end{aligned}$$

Therefore,

$$\begin{aligned}|E(v)-E(u)| &\le \int_\Omega \left[1+(|\nabla u|+|\nabla v|)^2\right]^{(p-1)/2}|\nabla(v-u)|\,dx \\ &\quad + \left|\int_\Omega f\,(v-u)\,dx\right| \\ &\le c\left(1+\|u\|_V^{p-1}+\|v\|_V^{p-1}+\|f\|_{V'}\right)\|v-u\|_V,\end{aligned}$$

from which, continuity of $E(\cdot)$ follows. □

Lemma 8.8.3 *The energy functional $E(\cdot)$ is strictly convex.*

Proof. This follows immediately from the strict convexity of the real-valued function $\boldsymbol{\xi} \mapsto \frac{1}{p}(1+|\boldsymbol{\xi}|^2)^{p/2}$ (see Exercise 5.3.13). □

Lemma 8.8.4 *The energy functional $E(\cdot)$ is Gâteaux differentiable and*

$$\langle E'(u), v\rangle = \int_\Omega (1+|\nabla u|^2)^{p/2-1}\nabla u\cdot\nabla v\,dx - \int_\Omega f\,v\,dx, \quad u,v\in V. \tag{8.8.11}$$

Proof. Similar to the proof of Lemma 8.8.2, we write

$$\begin{aligned}&\frac{1}{t}\left[E(u+tv)-E(u)\right] \\ &\quad = \int_\Omega\int_0^1 (1+|\nabla u+\tau\,t\,\nabla v|^2)^{p/2-1}(\nabla u+\tau\,t\,\nabla v)\cdot\nabla v\,d\tau\,dx - \int_\Omega f\,v\,dx\end{aligned}$$

for $u, v \in V$ and $t \neq 0$. Let $0 < |t| < 1$. Then

$$\begin{aligned}&\left|\int_0^1 (1+|\nabla u+\tau\,t\,\nabla v|^2)^{p/2-1}(\nabla u+\tau\,t\,\nabla v)\cdot\nabla v\,d\tau\right| \\ &\qquad \le \left[1+(|\nabla u|+|\nabla v|)^2\right]^{(p-1)/2}|\nabla v|.\end{aligned}$$

The right hand side is in $L^1(\Omega)$ by the Hölder inequality. So applying the Lebesgue Dominated Convergence Theorem 1.2.26, we know that $E(\cdot)$ is Gâteaux differentiable and we have the formula (8.8.11). □

We can now state the main result concerning the existence and uniqueness for the weak formulation (8.8.6) and the minimization problem (8.8.9).

Theorem 8.8.5 *Assume $f \in V'$ and $p \in (1,\infty)$. Then the weak formulation* (8.8.6) *and the minimization problem* (8.8.9) *are equivalent, and both admit a unique solution.*

Proof. Since V is reflexive, $E : V \to \mathbb{R}$ is coercive, continuous and strictly convex, by Theorem 3.3.12, we conclude the minimization problem (8.8.9) has a unique minimizer $u \in V$. By Theorem 5.3.19, we know that the weak formulation (8.8.6) and the minimization problem (8.8.9) are equivalent. □

Exercise 8.8.1 Use Theorem 7.3.13 to show that (8.8.5) defines a norm over the space V, which is equivalent to the standard norm $\|\cdot\|_{W^{1,p}(\Omega)}$.

Exercise 8.8.2 Apply Hölder's inequality (Lemma 1.5.3) to show that the left side of (8.8.3) makes sense for $u, v \in W^{1,p}(\Omega)$.

Exercise 8.8.3 For $p \geq 2$, show that the functional $E(\cdot)$ is Fréchet differentiable, and the formula (8.8.11) holds.

Exercise 8.8.4 For $p \in (1, \infty)$, consider the Neumann boundary value problem

$$-\operatorname{div}\left[(1+|\nabla u|^2)^{p/2-1}\nabla u\right] + b\,u = f \quad \text{in } \Omega,$$
$$(1+|\nabla u|^2)^{p/2-1}\frac{\partial u}{\partial \nu} = g \quad \text{on } \Gamma.$$

Make suitable assumptions on the data b, f, and g, and prove a result similar to Theorem 8.8.5.

Suggestion for Further Reading.

Many books can be consulted on detailed treatment of PDEs, for both steady and evolution equations, e.g., Evans [78], Lions and Magenes [158], McOwen [168], Wloka [233].

9
The Galerkin Method and Its Variants

In this chapter, we briefly discuss some numerical methods for solving boundary value problems. These are the Galerkin method and its variants: the Petrov-Galerkin method and the generalized Galerkin method. In Section 9.4, we rephrase the conjugate gradient method, discussed in Section 5.6, for solving variational equations.

9.1 The Galerkin method

The Galerkin method provides a general framework for approximation of operator equations, which includes the finite element method as a special case. In this section, we discuss the Galerkin method for a linear operator equation in a form directly applicable to the study of the finite element method.

Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bilinear form, and $\ell \in V'$. We consider the problem

$$u \in V, \quad a(u,v) = \ell(v) \quad \forall\, v \in V. \tag{9.1.1}$$

Throughout this section, we assume $a(\cdot,\cdot)$ is bounded,

$$|a(u,v)| \le M\, \|u\|_V \|v\|_V \quad \forall\, u, v \in V, \tag{9.1.2}$$

and V-elliptic,

$$a(v,v) \ge c_0 \|v\|_V^2 \quad \forall\, v \in V \tag{9.1.3}$$

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_9,
© Springer Science + Business Media, LLC 2009

for some positive constants M and c_0. Then according to the Lax-Milgram Lemma (Theorem 8.3.4), the variational problem (9.1.1) has a unique solution.

In general, it is impossible to find the exact solution of the problem (9.1.1) because the space V is infinite dimensional. A natural approach to constructing an approximate solution is to solve a finite dimensional analog of the problem (9.1.1). Thus, let $V_N \subset V$ be an N-dimensional subspace. We project the problem (9.1.1) onto V_N,

$$u_N \in V_N, \quad a(u_N, v) = \ell(v) \quad \forall\, v \in V_N. \tag{9.1.4}$$

Under the assumptions that the bilinear form $a(\cdot,\cdot)$ is bounded and V-elliptic, and $\ell \in V'$, we can again apply the Lax-Milgram Lemma and conclude that the problem (9.1.4) has a unique solution u_N.

We can express the problem (9.1.4) in the form of a linear system. Indeed, let $\{\phi_i\}_{i=1}^N$ be a basis of the finite dimensional space V_N. We write

$$u_N = \sum_{j=1}^{N} \xi_j \phi_j$$

and take $v \in V_N$ in (9.1.4) to be each of the basis functions ϕ_i. It is readily seen that (9.1.4) is equivalent to a linear system

$$A\boldsymbol{\xi} = \boldsymbol{b}. \tag{9.1.5}$$

Here, $\boldsymbol{\xi} = (\xi_j) \in \mathbb{R}^N$ is the unknown vector, $A = (a(\phi_j, \phi_i)) \in \mathbb{R}^{N\times N}$ is called the *stiffness matrix*, $\boldsymbol{b} = (\ell(\phi_i)) \in \mathbb{R}^N$ is the *load vector*. So the solution of the problem (9.1.4) can be found by solving a linear system.

The approximate solution u_N is, in general, different from the exact solution u. To increase the accuracy, it is natural to seek the approximate solution u_N in a larger subspace V_N. Thus, for a sequence of subspaces $V_{N_1} \subset V_{N_2} \subset \cdots \subset V$, we compute a corresponding sequence of approximate solutions $u_{N_i} \in V_{N_i}$, $i = 1, 2, \ldots$. This solution procedure is called the *Galerkin method*.

In the special case where additionally, the bilinear form $a(\cdot,\cdot)$ is symmetric,

$$a(u, v) = a(v, u) \quad \forall\, u, v \in V,$$

the original problem (9.1.1) is equivalent to a minimization problem

$$u \in V, \quad E(u) = \inf_{v\in V} E(v), \tag{9.1.6}$$

where the energy functional

$$E(v) = \frac{1}{2}\, a(v, v) - \ell(v). \tag{9.1.7}$$

With a finite dimensional subspace $V_N \subset V$ chosen, it is equally natural to develop a numerical method by minimizing the energy functional over the finite dimensional space V_N:

$$u_N \in V_N, \quad E(u_N) = \inf_{v \in V_N} E(v). \tag{9.1.8}$$

It is easy to verify that the two approximate problems (9.1.4) and (9.1.8) are equivalent. The method based on minimizing the energy functional over finite dimensional subspaces is called the *Ritz method*. From the above discussion, we see that the Galerkin method is more general than the Ritz method, and when both methods are applicable, they are equivalent. Usually, the Galerkin method is also called the *Ritz-Galerkin method*.

Example 9.1.1 We examine a concrete example of the Galerkin method. Consider the boundary value problem

$$\begin{cases} -u'' = f \quad \text{in } (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{9.1.9}$$

The weak formulation of the problem is

$$u \in V, \quad \int_0^1 u'v'dx = \int_0^1 fv\,dx \quad \forall\, v \in V,$$

where $V = H_0^1(0,1)$. Applying the Lax-Milgram Lemma, we see that the weak problem has a unique solution. To develop a Galerkin method, we need to choose a finite-dimensional subspace of V. Notice that a function in V must vanish at both $x = 0$ and $x = 1$. Thus a natural choice is

$$V_N = \operatorname{span}\left\{x^i(1-x) \mid i = 1, \ldots, N\right\}.$$

We write

$$u_N(x) = \sum_{j=1}^{N} \xi_j x^j (1-x);$$

the coefficients $\{\xi_j\}_{j=1}^N$ are determined by the Galerkin equations

$$\int_0^1 u_N' v' dx = \int_0^1 fv\,dx \quad \forall\, v \in V_N.$$

Taking v to be each of the basis functions $x^i(1-x)$, $1 \le i \le N$, we derive a linear system for the coefficients:

$$A\boldsymbol{\xi} = \boldsymbol{b}.$$

N	Cond (A)
3	8.92E+02
4	2.42E+04
5	6.56E+05
6	1.79E+07
7	4.95E+08
8	1.39E+10
9	3.93E+11
10	1.14E+13

TABLE 9.1. Condition numbers for the matrix in Example 9.1.1

Here $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^T$ is the vector of unknowns, $\boldsymbol{b} \in \mathbb{R}^N$ is a vector whose i^{th} component is $\int_0^1 f(x)\, x^i(1-x)\, dx$. The coefficient matrix is A, whose (i,j)-th entry is

$$\int_0^1 [x^j(1-x)]'[x^i(1-x)]' dx = \frac{(i+1)\,(j+1)}{i+j+1} + \frac{(i+2)\,(j+2)}{i+j+3} - \frac{(i+1)\,(j+2)+(i+2)\,(j+1)}{i+j+2}.$$

The coefficient matrix is rather ill-conditioned, indicating that it is difficult to solve the above Galerkin system numerically. Table 9.1 shows how rapidly the condition number of the matrix (measured in 2-norm) increases with the order N. We conclude that the seemingly natural choice of the basis functions $\{x^i(1-x)\}$ is not suitable in using the Galerkin method to solve the problem (9.1.9). □

Example 9.1.2 Let us consider the problem (9.1.9) again. This time, the finite-dimensional subspace is chosen to be

$$V_N = \text{span}\,\{\sin(i\pi x) \mid i = 1, \ldots, N\}.$$

The basis functions are orthogonal with respect to the inner product defined by the bilinear form:

$$\int_0^1 (\sin j\pi x)'(\sin i\pi x)' dx = ij\pi^2 \int_0^1 \cos j\pi x \, \cos i\pi x \, dx = \frac{ij\pi^2}{2}\, \delta_{ij}.$$

Writing

$$u_N(x) = \sum_{j=1}^{N} \xi_j \sin j\pi x,$$

we see that the coefficients $\{\xi_j\}_{j=1}^N$ are determined by the linear system

$$\sum_{j=1}^{N} \xi_j \int_0^1 (\sin j\pi x)'(\sin i\pi x)' dx = \int_0^1 f(x)\, \sin i\pi x\, dx, \quad i = 1, \ldots, N.$$

This is a diagonal system and we easily find the solution:

$$\xi_i = \frac{2}{\pi^2 i^2} \int_0^1 f(x)\, \sin i\pi x\, dx, \quad i = 1, \ldots, N.$$

It is worth noticing that the Galerkin solution can be written in the form of a kernel approximation:

$$u_N(x) = \int_0^1 f(t)\, K_N(x,t)\, dt, \tag{9.1.10}$$

where the kernel function

$$K_N(x,t) = \frac{2}{\pi^2} \sum_{j=1}^{N} \frac{\sin j\pi x\, \sin j\pi t}{j^2}. \qquad \square$$

From the above two examples, we see that in applying the Galerkin method it is very important to choose appropriate basis functions for finite-dimensional subspaces. Before the invention of computers, the Galerkin method was applied mainly with the use of global polynomials or global trigonometric polynomials. For the simple model problem (9.1.9) we see that the simpll choice of the polynomial basis functions $\{x^i(1-x)\}$ leads to a severely ill-conditioned linear system. For the same model problem, the trigonometric polynomial basis functions $\{\sin(i\pi x)\}$ is ideal in the sense that it leads to a diagonal linear system so that its conditioning is best possible. We need to be aware, though, that trigonometric polynomial basis functions can lead to severely ill-conditioned linear systems in different but equally simple model problems. The idea of the finite element method (see Chapter 10) is to use basis functions with small supports so that, among various advantages of the method, the conditioning of the resulting linear system can be moderately maintained (see Table 10.1 and Exercise 10.3.7 for an estimate on the growth of the condition number of stiffness matrices as the mesh is refined).

Now we consider the important issue of convergence and error estimation for the Galerkin method. A key result is the following Céa's inequality.

Proposition 9.1.3 *Assume V is a Hilbert space, $V_N \subset V$ is a subspace, $a(\cdot,\cdot)$ is a bounded, V-elliptic bilinear form on V, and $\ell \in V'$. Let $u \in V$ be the solution of the problem* (9.1.1)*, and $u_N \in V_N$ be the Galerkin approximation defined in* (9.1.4)*. Then there is a constant c such that*

$$\|u - u_N\|_V \le c \inf_{v \in V_N} \|u - v\|_V. \tag{9.1.11}$$

Proof. Subtracting (9.1.4) from (9.1.1) with $v \in V_N$, we obtain an error relation

$$a(u - u_N, v) = 0 \quad \forall\, v \in V_N. \tag{9.1.12}$$

Using the V-ellipticity of $a(\cdot,\cdot)$, the error relation and the boundedness of $a(\cdot,\cdot)$, we have, for any $v \in V_N$,

$$\begin{aligned} c_0\|u - u_N\|_V^2 &\le a(u - u_N, u - u_N) \\ &= a(u - u_N, u - v) \\ &\le M\,\|u - u_N\|_V \|u - v\|_V. \end{aligned}$$

Thus

$$\|u - u_N\|_V \le c\,\|u - v\|_V,$$

where we may take $c = M/c_0$. Since v is arbitrary in V_N, we have the inequality (9.1.11). □

The inequality (9.1.11) is known as Céa's lemma in the literature. Such an inequality was first proved by Céa [47] for the case when the bilinear form is symmetric and was extended to the non-symmetric case in [37]. The inequality (9.1.11) states that to estimate the error of the Galerkin solution, it suffices to estimate the approximation error $\inf_{v\in V_N} \|u - v\|$.

In the special case where $a(\cdot,\cdot)$ is also symmetric, we may assign a geometrical interpretation of the error relation (9.1.12). Indeed, in this special case, the bilinear form $a(\cdot,\cdot)$ defines an inner product over the space V and its induced norm $\|v\|_a = \sqrt{a(v,v)}$, called the energy norm, is equivalent to the norm $\|v\|_V$. With respect to this new inner product, the Galerkin solution error $u - u_N$ is orthogonal to the subspace V_N, or in other words, the Galerkin solution u_N is the orthogonal projection of the exact solution u to the subspace V_N. Also in this special case, Céa's inequality (9.1.11) can be replaced by

$$\|u - u_N\|_a = \inf_{v\in V_N} \|u - v\|_a,$$

i.e., measured in the energy norm, u_N is the optimal approximation of u from the subspace V_N.

Céa's inequality is a basis for convergence analysis and error estimations. As a simple consequence, we have the next convergence result.

Corollary 9.1.4 *We make the assumptions stated in Proposition 9.1.3. Assume $V_1 \subset V_2 \subset \cdots$ is a sequence of finite dimensional subspaces of V with the property*

$$\overline{\bigcup_{n\ge 1} V_n} = V. \tag{9.1.13}$$

Then the Galerkin method converges:

$$\|u - u_n\|_V \to 0 \quad \text{as } n \to \infty, \tag{9.1.14}$$

where $u_n \in V_n$ is the Galerkin solution defined by (9.1.4).

Proof. By the density assumption (9.1.13), we can find a sequence $v_n \in V_n$, $n \geq 1$, such that

$$\|u - v_n\|_V \to 0 \quad \text{as } n \to \infty.$$

Applying Céa's inequality (9.1.11), we have

$$\|u - u_n\|_V \leq c \, \|u - v_n\|_V.$$

Therefore, we have the convergence statement (9.1.14). □

Note that the assumptions made on the subspace sequence $\{V_n\}_{n\geq 1}$ requires V to be a separable Hilbert space. This is not a restriction in applications for solving boundary value problems, since Sobolev spaces $H^k(\Omega)$ and their subspaces are separable.

When the finite dimensional space V_N is constructed from piecewise (images of) polynomials, the Galerkin method leads to a finite element method, which will be discussed in some detail in the chapter following. We will see in the context of the finite element method that Céa's inequality also serves as a basis for error estimates.

Exercise 9.1.1 Show that the discrete problems (9.1.4) and (9.1.8) are equivalent.

Exercise 9.1.2 Show that if the bilinear form $a(\cdot, \cdot)$ is symmetric, then the stiffness matrix A is symmetric; if $a(\cdot, \cdot)$ is V-elliptic, then A is positive definite.

Exercise 9.1.3 From Exercise 2.3.9, we know the solution of the problem (9.1.9) is:

$$u(x) = \int_0^1 f(t) \, K(x, t) \, dt, \tag{9.1.15}$$

where

$$K(x, t) = \min(x, t) \, (1 - \max(x, t)) \, .$$

Show that the kernel function has the Fourier expansion

$$K(x, t) = \frac{2}{\pi^2} \sum_{j=1}^{\infty} \frac{\sin j\pi x \, \sin j\pi t}{j^2}.$$

Thus, the Galerkin solution (9.1.10) can be viewed as being obtained from (9.1.15) by truncating the Fourier series of the kernel function.

Exercise 9.1.4 The Galerkin method is not just a general framework for developing numerical schemes; it can be used to provide another proof of the existence of a solution to a problem like (9.1.1). We make the assumptions stated in Corollary 9.1.4. Fill details of the following steps.
(a) Prove the unique solvability of the finite-dimensional linear system (9.1.4) by showing zero is the only solution of the system corresponding to $\ell = 0$.
(b) Show that the sequence $\{u_n\}_{n\geq 1}$ is bounded, and thus it has a subsequence $\{u_{n_i}\}_{i\geq 1}$ converging weakly to some element $u \in V$.

(c) For any fixed positive integer N and $v \in V_N$, show that $a(u, v) = \ell(v)$.
(d) Use the condition (9.1.13) to show that (9.1.1) is valid.
(e) Show the entire sequence $\{u_n\}_{n\geq 1}$ converges strongly to u.

9.2 The Petrov-Galerkin method

The Petrov-Galerkin method for a linear boundary value problem can be developed based on the framework of the generalized Lax-Milgram Lemma presented in Section 8.7. Let U and V be two real Hilbert spaces, $a : U \times V \to \mathbb{R}$ be a bilinear form, and $\ell \in V'$. The problem to be solved is

$$u \in U, \quad a(u, v) = \ell(v) \quad \forall\, v \in V. \tag{9.2.1}$$

From the generalized Lax-Milgram Lemma, we know that the problem (9.2.1) has a unique solution $u \in U$, if the following conditions are satisfied: there exist constants $M > 0$ and $\alpha > 0$ such that

$$|a(u, v)| \leq M \|u\|_U \|v\|_V \quad \forall\, u \in U,\ v \in V, \tag{9.2.2}$$

$$\sup_{0\neq v\in V} \frac{a(u, v)}{\|v\|_V} \geq \alpha \|u\|_U \quad \forall\, u \in U, \tag{9.2.3}$$

$$\sup_{u\in U} a(u, v) > 0 \quad \forall\, v \in V,\ v \neq 0. \tag{9.2.4}$$

Now let $U_N \subset U$ and $V_N \subset V$ be finite dimensional subspaces of U and V with $\dim(U_N) = \dim(V_N) = N$. Then a Petrov-Galerkin method to solve the problem (9.2.1) is given by

$$u_N \in U_N, \quad a(u_N, v_N) = \ell(v_N) \quad \forall\, v_N \in V_N. \tag{9.2.5}$$

Well-posedness and error analysis for the method (9.2.5) are discussed in the next result (see [29]).

Theorem 9.2.1 *We keep the above assumptions on the spaces U, V, U_N and V_N, and on the forms $a(\cdot, \cdot)$ and $\ell(\cdot)$. Assume further that there exists a constant $\alpha_N > 0$, such that*

$$\sup_{0\neq v_N\in V_N} \frac{a(u_N, v_N)}{\|v_N\|_V} \geq \alpha_N \|u_N\|_U \quad \forall\, u_N \in U_N. \tag{9.2.6}$$

Then the discrete problem (9.2.5) *has a unique solution u_N, and we have the error estimate*

$$\|u - u_N\|_U \leq \left(1 + \frac{M}{\alpha_N}\right) \inf_{w_N\in U_N} \|u - w_N\|_U. \tag{9.2.7}$$

Proof. The assumption (9.2.6) implies that the only solution of the homogeneous of the problem (9.2.5) is $u_N = 0$. Hence, the problem (9.2.5) has a unique solution u_N. Subtracting (9.2.5) from (9.2.1) with $v = v_N \in V_N$, we obtain the error relation

$$a(u - u_N, v_N) = 0 \quad \forall\, v_N \in V_N. \tag{9.2.8}$$

Now for any $w_N \in V_N$, we write

$$\|u - u_N\|_U \le \|u - w_N\|_U + \|u_N - w_N\|_U. \tag{9.2.9}$$

Using the condition (9.2.6), we have

$$\alpha_N \|u_N - w_N\|_U \le \sup_{0 \neq v_N \in V_N} \frac{a(u_N - w_N, v_N)}{\|v_N\|_V}.$$

Using the error relation (9.2.8), we then obtain

$$\alpha_N \|u_N - w_N\|_U \le \sup_{0 \neq v_N \in V_N} \frac{a(u - w_N, v_N)}{\|v_N\|_V}.$$

The right hand side can be bounded by $M\,\|u - w_N\|_U$. Therefore,

$$\|u_N - w_N\|_U \le \frac{M}{\alpha_N}\,\|u - w_N\|_U.$$

This inequality and (9.2.9) imply the estimate (9.2.7). □

Remark 9.2.2 The error bound (9.2.7) is improved to

$$\|u - u_N\|_U \le \frac{M}{\alpha_N} \inf_{w_N \in U_N} \|u - w_N\|_U \tag{9.2.10}$$

in [236]. In the case of the Galerkin method, $U = V$ and $U_N = V_N$; the error bound (9.2.10) reduces to Céa's inequality (9.1.11); recall from the proof of Proposition 9.1.3 that the constant c in the inequality (9.1.11) is M/α_N in the notation adopted here. An essential ingredient in the proof of (9.2.10) in [236] is the following result concerning a projection operator:

> Let H be a Hilbert space, and $P : H \to H$ a linear operator satisfying $0 \neq P^2 = P \neq I$. Then
>
> $$\|P\|_{H,H} = \|I - P\|_{H,H}.$$

This result is applied to the linear operator $P_N : U \to U$ defined by $P_N u = u_N$. □

As in Corollary 9.1.4, we have a convergence result based on the estimate (9.2.7).

Corollary 9.2.3 *We make the assumptions stated in Theorem 9.2.1. Furthermore, we assume that there is a constant $\alpha_0 > 0$ such that*

$$\alpha_N \geq \alpha_0 \quad \forall\, N. \tag{9.2.11}$$

Assume the sequence of subspaces $U_{N_1} \subset U_{N_2} \subset \cdots \subset U$ has the property

$$\overline{\bigcup_{i\geq 1} U_{N_i}} = U. \tag{9.2.12}$$

Then the Petrov-Galerkin method (9.2.5) *converges:*

$$\|u - u_{N_i}\|_U \to 0 \quad \text{as } i \to \infty.$$

We remark that to achieve convergence of the method, we can allow α_N to approach 0 under certain rule, as long as

$$\max\{1, \alpha_N^{-1}\} \inf_{w_N \in U_N} \|u - w_N\|_U \to 0$$

as is seen from the bound (9.2.7). Nevertheless, the condition (9.2.11) is crucial in obtaining optimal order error estimates. This condition is usually written as

$$\sup_{0\neq v_N \in V_N} \frac{a(u_N, v_N)}{\|v_N\|_V} \geq \alpha_0 \|u_N\|_U \quad \forall\, u_N \in U_N,\ \forall\, N, \tag{9.2.13}$$

or equivalently,

$$\inf_{0\neq u_N \in U_N} \sup_{0\neq v_N \in V_N} \frac{a(u_N, v_N)}{\|u_N\|_U \|v_N\|_V} \geq \alpha_0 \quad \forall\, N. \tag{9.2.14}$$

In the literature, this condition is called the inf-sup condition or Babuška-Brezzi condition. This condition states that the two finite dimensional spaces must be compatible in order to yield convergent numerical solutions. The condition is most important in the context of the study of mixed finite element methods ([43]).

9.3 Generalized Galerkin method

In the Galerkin method discussed in Section 9.1, the finite dimensional space V_N is assumed to be a subspace of V. The resulting numerical method is called an internal approximation method. For certain problems, we will need to relax this assumption and to allow the variational "crime" $V_N \not\subset$

V. This, for instance, is the case for non-conforming method. There are situations where considerations of other variational "crimes" are needed. Two such situations are where a general curved domain is approximated by a polyhedral/polygonal domain and where numerical quadratures are used to compute the integrals defining the bilinear form and the linear form. These considerations lead to the following framework of a generalized Galerkin method for the approximate solution of the problem (9.1.1):

$$u_N \in V_N, \quad a_N(u_N, v_N) = \ell_N(v_N) \quad \forall\, v_N \in V_N. \tag{9.3.1}$$

Here, V_N is a finite dimensional space, but it is no longer assumed to be a subspace of V; the bilinear form $a_N(\cdot,\cdot)$ and the linear form $\ell_N(\cdot)$ are suitable approximations of $a(\cdot,\cdot)$ and $\ell(\cdot)$.

We have the following result related to the approximation method (9.3.1).

Theorem 9.3.1 *Assume a discretization dependent norm $\|\cdot\|_N$, the approximation bilinear form $a_N(\cdot,\cdot)$ and the linear form $\ell_N(\cdot)$ are defined on the space*

$$V + V_N = \{w \mid w = v + v_N,\ v \in V,\ v_N \in V_N\}.$$

Assume there exist constants $M, \alpha_0, c_0 > 0$, independent of N, such that

$$\begin{aligned} |a_N(w, v_N)| &\le M\, \|w\|_N \|v_N\|_N \quad \forall\, w \in V + V_N,\ \forall\, v_N \in V_N, \\ a_N(v_N, v_N) &\ge \alpha_0 \|v_N\|_N^2 \quad \forall\, v_N \in V_N, \\ |\ell_N(v_N)| &\le c_0 \|v_N\|_N \quad \forall\, v_N \in V_N. \end{aligned}$$

Then the problem (9.3.1) *has a unique solution $u_N \in V_N$, and we have the error estimate*

$$\begin{aligned} \|u - u_N\|_N \le \left(1 + \frac{M}{\alpha_0}\right) & \inf_{w_N \in V_N} \|u - w_N\|_N \\ & + \frac{1}{\alpha_0} \sup_{v_N \in V_N} \frac{|a_N(u, v_N) - \ell_N(v_N)|}{\|v_N\|_N}. \end{aligned} \tag{9.3.2}$$

Proof. The unique solvability of the problem (9.3.1) follows from an application of the Lax-Milgram Lemma. Let us derive the error estimate (9.3.2). For any $w_N \in V_N$, we write

$$\|u - u_N\|_N \le \|u - w_N\|_N + \|w_N - u_N\|_N.$$

Using the assumptions on the approximate bilinear form and the definition of the approximate solution u_N, we have

$$\begin{aligned} &\alpha_0 \|w_N - u_N\|_N^2 \\ &\quad \le a_N(w_N - u_N, w_N - u_N) \\ &\quad = a_N(w_N - u, w_N - u_N) + a_N(u, w_N - u_N) - \ell_N(w_N - u_N) \\ &\quad \le M\, \|w_N - u\|_N \|w_N - u_N\|_N + |a_N(u, w_N - u_N) - \ell_N(w_N - u_N)|. \end{aligned}$$

Thus

$$\alpha_0 \|w_N - u_N\|_N \leq M \|w_N - u\|_N + \frac{|a_N(u, w_N - u_N) - \ell_N(w_N - u_N)|}{\|w_N - u_N\|_N}.$$

We replace $w_N - u_N$ by v_N and take the supremum of the second term of the right hand side with respect to $v_N \in V_N$ to obtain (9.3.2). □

The estimate (9.3.2) is a Strang type estimate for the effect of the variational "crimes" on the numerical solution. We notice that in the bound of the estimate (9.3.2), the first term is on the approximation property of the solution u by functions from the finite dimensional space V_N, whereas the second term describes the extent to which the exact solution u satisfies the approximate problem.

Exercise 9.3.1 Show that in the case of a conforming method (i.e., $V_N \subset V$. $a_N(\cdot, \cdot) \equiv a(\cdot, \cdot)$ and $\ell_N(\cdot) \equiv \ell(\cdot)$), the error estimate (9.3.2) reduces to an inequality of the form (9.1.11).

9.4 Conjugate gradient method: variational formulation

In Section 5.6, we discussed the conjugate gradient method for a linear operator equation. In this section, we present a form of the method for variational equations. For this purpose, let V be a real Hilbert space with inner product $(\cdot, \cdot)$ and norm $\|\cdot\|$, and let $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a continuous, symmetric, V-elliptic bilinear form. Recall that the continuity assumption implies the existence of a constant $M > 0$ such that

$$|a(u, v)| \leq M \|u\| \|v\| \quad \forall\, u, v \in V, \tag{9.4.1}$$

the symmetry assumption is

$$a(u, v) = a(v, u) \quad \forall\, u, v \in V, \tag{9.4.2}$$

and the V-ellipticity assumption implies the existence of a constant $\alpha > 0$ such that

$$a(v, v) \geq \alpha \|v\|^2 \quad \forall\, v \in V. \tag{9.4.3}$$

Given $\ell \in V'$, the variational problem is

$$u \in V, \quad a(u, v) = \ell(v) \quad \forall\, v \in V. \tag{9.4.4}$$

Under the stated assumptions, by the Lax-Milgram Lemma, the problem (9.4.4) has a unique solution $u \in V$.

We turn to the conjugate gradient method for solving the variational problem (9.4.4). For this, we first rewrite it as an operator equation of the form $Au = f$. Introduce an operator $A : V \to V$ and an element $f \in V$ defined by

$$(Au, v) = a(u, v) \quad \forall\, u, v \in V, \tag{9.4.5}$$

$$(f, v) = \ell(v) \quad \forall\, v \in V. \tag{9.4.6}$$

The existence of A and f is ensured by the Riesz representation theorem; see also the beginning of Section 8.3. From the assumptions on the bilinear form $a(\cdot, \cdot)$, we know that A is bounded, self-adjoint, and positive definite:

$$\begin{aligned} \|A\| &\le M, \\ (Au, v) &= (u, Av) \quad \forall\, u, v \in V, \\ (Av, v) &\ge \alpha \|v\|^2 \quad \forall\, v \in V. \end{aligned}$$

Also, $\|f\| = \|\ell\|$ between the V-norm of f and V'-norm of ℓ. With A and f given by (9.4.5) and (9.4.6), the problem (9.4.4) is equivalent to the operator equation

$$Au = f. \tag{9.4.7}$$

Now the conjugate gradient method presented in Section 5.6 can be stated for the problem (9.4.4) as follows.

Algorithm 1. Conjugate gradient method for the linear problem (9.4.4):

Initialization.

Choose an initial guess $u_0 \in V$.

Compute the initial residual

$$r_0 \in V, \quad (r_0, v) = \ell(v) - a(u_0, v) \quad \forall\, v \in V.$$

Define the descent direction $s_0 = r_0$.

Iteration.

For $k = 0, 1, \cdots$, compute the scalar

$$\alpha_k = \frac{\|r_k\|^2}{a(s_k, s_k)},$$

define the new iterate

$$u_{k+1} = u_k + \alpha_k s_k,$$

and compute its residual

$$r_{k+1} \in V, \quad (r_{k+1}, v) = \ell(v) - a(u_{k+1}, v) \quad \forall\, v \in V.$$

Then compute the scalar

$$\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$$

and define the new descent direction

$$s_{k+1} = r_{k+1} + \beta_k s_k.$$

Convergence of Algorithm 1 follows from Theorem 5.6.1. As one possible stopping criterion, we may test the convergence condition $\|r_{k+1}\| \le \varepsilon$ for a given small threshold $\varepsilon > 0$.

We now extend the conjugate gradient method for some nonlinear problems. Observe that under the stated assumptions on $a(\cdot,\cdot)$ and ℓ, the linear problem (9.4.4) is equivalent to a minimization problem for a quadratic functional:

$$u \in V, \quad E(u) = \inf_{v \in V} E(v),$$

where the "energy" functional

$$E(v) = \frac{1}{2}\, a(v,v) - \ell(v).$$

Note that the Gâteaux derivative $E' : V \to V'$ is given by

$$\langle E'(u), v\rangle = a(u,v) - \ell(v),$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between V' and V. Thus the residual equations in Algorithm 1 can be put in a different form: for $k = 0, 1, \cdots$,

$$r_k \in V, \quad (r_k, v) = -\langle E'(u_k), v\rangle \quad \forall\, v \in V.$$

Based on this observation, we can introduce a conjugate gradient algorithm for solving a general optimization problem

$$u \in V, \quad J(u) = \inf_{v \in V} J(v). \tag{9.4.8}$$

We assume the functional $J : V \to \mathbb{R}$ is strictly convex and Gâteaux differentiable. Then from the theory in Section 3.3, we know that the problem (9.4.8) has a unique solution. Motivated by Algorithm 1, we have the following conjugate gradient method for (9.4.8).

Algorithm 2. Conjugate gradient method for the optimization problem (9.4.8):

Initialization.

Choose an initial guess $u_0 \in V$.

Compute the initial residual

$$r_0 \in V, \quad (r_0, v) = -\langle J'(u_0), v\rangle \quad \forall\, v \in V.$$

Define the descent direction $s_0 = r_0$.

Iteration.

For $k = 0, 1, \cdots$, define the new iterate

$$u_{k+1} = u_k + \alpha_k s_k,$$

where $\alpha_k \in \mathbb{R}$ is a solution of the one-dimensional problem

$$J(u_k + \alpha_k s_k) = \inf_{\alpha \in \mathbb{R}} J(u_k + \alpha\, s_k),$$

and compute the residual

$$r_{k+1} \in V, \quad (r_{k+1}, v) = -\langle J'(u_{k+1}), v\rangle \quad \forall\, v \in V.$$

Then compute the scalar

$$\beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}$$

and define the new descent direction

$$s_{k+1} = r_{k+1} + \beta_k s_k.$$

This algorithm converges if $J \in C^1(V;\mathbb{R})$ is coercive and J' is Lipschitz continuous and strongly monotone on any bounded set in V ([92]). The assumption on J' means that for any $R > 0$, there are constants $M_R, \alpha_R > 0$, depending on R, such that for any $u, v \in V$ with $\|u\| \le R$ and $\|v\| \le R$,

$$\|J'(u) - J'(v)\|_{V'} \le M_R \|u - v\|_V,$$
$$\langle J'(u) - J'(v), u - v\rangle \ge \alpha_R \|u - v\|_V^2.$$

Suggestion for Further Reading.

Based on any weak formulation, we can develop a particular Galerkin type numerical method. Mixed formulations are the basis for mixed Galerkin finite element methods. We refer the reader to [43] for an extensive treatment of the mixed methods.

Many numerical methods exist for solving differential equations. In this text, we do not touch upon some other popular methods, e.g., the collocation method, the spectral method, the finite volume method, etc. and various combinations of these methods such as the spectral collocation method. The well-written book [189] can be consulted for discussions of

many of the existing methods for numerical solution of partial differential equations and for a rather comprehensive list of related references.

Some references on the conjugate gradient method, mainly for finite dimensional systems, are listed at the end of Chapter 5. Discussion of the method for infinite dimensional variational problems, especially those arising in computational fluid dynamics, can be found in [92].

10
Finite Element Analysis

The finite element method is the most popular numerical method for solving elliptic boundary value problems. In this chapter, we introduce the concept of the finite element method, the finite element interpolation theory and its application in error estimates of finite element solutions of elliptic boundary value problems. The boundary value problems considered in this chapter are linear.

From the discussion in the previous chapter, we see that the Galerkin method for a linear boundary value problem reduces to the solution of a linear system. In solving the linear system, properties of the coefficient matrix A play an essential role. For example, if the condition number of A is too big, then from a practical perspective, it is impossible to find directly an accurate solution of the system (see [15]). Another important issue is the sparsity of the matrix A. The matrix A is said to be *sparse*, if most of its entries are zero; otherwise the matrix is said to be *dense*. Sparseness of the matrix can be utilized for two purposes. First, the stiffness matrix is less costly to form (observing that the computation of each entry of the matrix involves a domain integration and sometimes a boundary integration as well). Second, if the coefficient matrix is sparse, then the linear system can usually be solved more efficiently. To get a sparse stiffness matrix with the Galerkin method, we use finite dimensional approximation spaces such that it is possible to choose basis functions with small support. This consideration gives rise to the idea of the finite element method, where we use piecewise (images of) smooth functions (usually polynomials) for approximations. Loosely speaking, the finite element method is a Galerkin method with the use of piecewise (images of) polynomials.

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_10,
© Springer Science + Business Media, LLC 2009

All the discussions made on the Galerkin method in Section 9.1 are valid for the finite element method. In particular, we still have Céa's inequality, and the problem of estimating the finite element solution error is reduced to one of estimating the approximation error

$$\|u - u_h\|_V \le c\,\|u - \Pi_h u\|_V,$$

where $\Pi_h u$ is a finite element interpolant of u. We will study in some detail affine families of finite elements and derive some order error estimates for finite element interpolants.

In Section 10.1, we examine some examples of the finite element method for solving one-dimensional boundary value problems. In Section 10.2, we discuss construction of finite element functions on a polygonal domain, with an emphasis on affine-equivalent families. In Section 10.3, we present interpolation error estimation for affine-equivalent finite elements. And finally, in Section 10.4, we consider convergence and error estimates for finite element solutions of elliptic boundary value problems.

10.1 One-dimensional examples

To have some idea of the finite element method, in this section we examine some examples on solving one-dimensional boundary value problems. These examples exhibit various aspects of the finite element method in the simple context of one-dimensional problems.

10.1.1 Linear elements for a second-order problem

Let us consider a finite element method to solve the boundary value problem

$$\begin{cases} -u'' + u = f \quad \text{in } \Omega = (0,1), \\ u(0) = 0,\ u'(1) = b, \end{cases} \tag{10.1.1}$$

where $f \in L^2(0,1)$ and $b \in \mathbb{R}$ are given. Let

$$V = H_{(0}(0,1) = \{v \in H^1(0,1) \mid v(0) = 0\},$$

a subspace of $H^1(0,1)$. The weak formulation of the problem is

$$u \in V, \quad \int_0^1 (u'v' + u\,v)\,dx = \int_0^1 f\,v\,dx + b\,v(1) \quad \forall\, v \in V. \tag{10.1.2}$$

Applying the Lax-Milgram Lemma, we see that the problem (10.1.2) has a unique solution.

Let us develop a finite element method for the problem. For a natural number N, we partition the set $\overline{\Omega} = [0,1]$ into N parts:

$$\overline{\Omega} = \cup_{i=1}^N K_i,$$

where $K_i = [x_{i-1}, x_i]$, $1 \le i \le N$, are called the elements, and the x_i, $0 \le i \le N$, are called the nodes, $0 = x_0 < x_1 < \cdots < x_N = 1$. In this example, we have a Dirichlet condition at the node x_0. Denote $h_i = x_i - x_{i-1}$, and $h = \max_{1 \le i \le N} h_i$. The value h is called the meshsize or mesh parameter. We use continuous piecewise linear functions for the approximation, i.e., we choose

$$V_h = \{v_h \in V \mid v_h|_{K_i} \in \mathbb{P}_1(K_i),\ 1 \le i \le N\}.$$

From the discussion in Chapter 7, we know that for a piecewisely smooth function v_h, $v_h \in H^1(\Omega)$ if and only if $v_h \in C(\overline{\Omega})$. Thus a more transparent yet equivalent definition of the finite element space is

$$V_h = \{v_h \in C(\overline{\Omega}) \mid v_h|_{K_i} \in \mathbb{P}_1(K_i),\ 1 \le i \le N,\ v_h(0) = 0\}.$$

For the basis functions of the space V_h, we introduce hat functions associated with the nodes $x_1, \ldots, x_N$. For $i = 1, \ldots, N-1$, let

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/h_i, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_{i+1}, & x_i \le x \le x_{i+1}, \\ 0, & \text{otherwise}, \end{cases} \tag{10.1.3}$$

and for $i = N$,

$$\phi_N(x) = \begin{cases} (x - x_{N-1})/h_N, & x_{N-1} \le x \le x_N, \\ 0, & \text{otherwise}. \end{cases} \tag{10.1.4}$$

These functions are continuous and piecewise linear (see Figure 10.1). It is easy to see they are linearly independent. The first order weak derivatives of the basis functions are piecewise constants. Indeed, for $i = 1, \ldots, N-1$,

$$\phi_i'(x) = \begin{cases} 1/h_i, & x_{i-1} < x < x_i, \\ -1/h_{i+1}, & x_i < x < x_{i+1}, \\ 0, & x < x_{i-1} \text{ or } x > x_{i+1}, \end{cases}$$

and for $i = N$,

$$\phi_N'(x) = \begin{cases} 1/h_N, & x_{N-1} < x \le x_N, \\ 0, & x < x_{N-1}. \end{cases}$$

Then the finite element space can be expressed as

$$V_h = \operatorname{span}\{\phi_i \mid 1 \le i \le N\},$$

i.e., any function in V_h is a linear combination of the hat functions $\{\phi_i\}_{i=1}^N$. The corresponding finite element method is

$$u_h \in V_h, \quad \int_0^1 (u_h' v_h' + u_h\, v_h)\, dx = \int_0^1 f\, v_h dx + b\, v_h(1) \quad \forall\, v_h \in V_h, \tag{10.1.5}$$

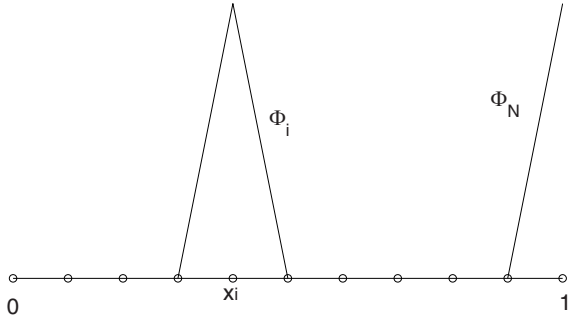

FIGURE 10.1. Piecewise linear basis functions

which admits a unique solution by another application of the Lax-Milgram Lemma. Write

$$u_h = \sum_{j=1}^{N} u_j \phi_j.$$

Note that $u_j = u_h(x_j)$, $1 \le j \le N$. We see that the finite element method (10.1.5) is equivalent to the following linear system for the unknowns $u_1, \ldots, u_N$:

$$\sum_{j=1}^{N} u_j \int_0^1 (\phi_i' \phi_j' + \phi_i \phi_j)\, dx = \int_0^1 f\, \phi_i dx + b\, \phi_i(1), \quad 1 \le i \le N. \quad (10.1.6)$$

Let us find the coefficient matrix of the system (10.1.6) in the case of a uniform partition, i.e., $h_1 = \cdots = h_N = h$. The following formulas are

useful for this purpose.

$$\begin{aligned}
\int_0^1 \phi_i' \phi_{i-1}' dx &= -\frac{1}{h}, \quad 2 \le i \le N, \\
\int_0^1 (\phi_i')^2 dx &= \frac{2}{h}, \quad 1 \le i \le N-1, \\
\int_0^1 \phi_i \phi_{i-1} dx &= \frac{h}{6}, \quad 2 \le i \le N, \\
\int_0^1 (\phi_i)^2 dx &= \frac{2\,h}{3}, \quad 1 \le i \le N-1, \\
\int_0^1 (\phi_N')^2 dx &= \frac{1}{h}, \\
\int_0^1 (\phi_N)^2 dx &= \frac{h}{3}.
\end{aligned}$$

We see that in matrix/vector notation, in the case of a uniform partition, the finite element system (10.1.6) can be written as

$$A\,\boldsymbol{u} = \boldsymbol{b},$$

where

$$\boldsymbol{u} = (u_1, \ldots, u_N)^T$$

is the unknown vector,

$$A = \begin{pmatrix}
\frac{2\,h}{3} + \frac{2}{h} & \frac{h}{6} - \frac{1}{h} & & & \\
\frac{h}{6} - \frac{1}{h} & \frac{2\,h}{3} + \frac{2}{h} & \frac{h}{6} - \frac{1}{h} & & \\
\ddots & \ddots & \ddots & & \\
 & \frac{h}{6} - \frac{1}{h} & \frac{2\,h}{3} + \frac{2}{h} & \frac{h}{6} - \frac{1}{h} & \\
 & & \frac{h}{6} - \frac{1}{h} & \frac{h}{3} + \frac{1}{h}
\end{pmatrix}_{N \times N} \tag{10.1.7}$$

is the stiffness matrix, and

$$\boldsymbol{b} = \left(\int_0^1 f\,\phi_1 dx, \ldots, \int_0^1 f\,\phi_{N-1} dx, \int_0^1 f\,\phi_N dx + b \right)^T$$

is the load vector. The matrix A is sparse, thanks to the small supports of the basis functions. One distinguished feature of the finite element method is that the basis functions are constructed in such a way that their supports are as small as possible, so that the corresponding stiffness matrix is as

sparse as possible. It can be shown that the condition number of the matrix A of (10.1.7) is $\mathcal{O}(h^{-2})$ (see Exercise 10.3.7).

Let us examine another example. Instead of the model problem (10.1.1), consider

$$\begin{cases} -u'' = f \quad \text{in } \Omega = (0,1), \\ u(0) = u(1) = 0, \end{cases} \tag{10.1.8}$$

where $f \in L^2(0,1)$ is given. The weak formulation is

$$u \in V, \quad \int_0^1 u'v'dx = \int_0^1 f\, v\, dx \quad \forall\, v \in V,$$

where $V = H_0^1(\Omega)$. Again, use the continuous piecewise linear functions associated with the uniform partition of the interval $[0,1]$. With the basis functions (10.1.3) where $h_i = h_{i+1} = h = 1/N$, the corresponding finite element space

$$V_h = \text{span}\,\{\phi_i \mid 1 \le i \le N-1\}.$$

Instead of (10.1.6), the linear system for the finite element solution

$$u_h = \sum_{j=1}^{N-1} u_j \phi_j$$

is now

$$\sum_{j=1}^{N-1} u_j \int_0^1 \phi_i' \phi_j' dx = \int_0^1 f\, \phi_i dx, \quad 1 \le i \le N-1.$$

Its coefficient matrix is

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}_{(N-1)\times(N-1)} . \tag{10.1.9}$$

Using the result from Exercise 6.3.1, we find

$$\text{Cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{1+\cos(\pi h)}{1-\cos(\pi h)} = \mathcal{O}(h^{-2}).$$

For h close to 0, we have

$$\text{Cond}_2(A) = \frac{4}{\pi^2} h^{-2} + \mathcal{O}(1).$$

We list in Table 10.1 values of $\text{Cond}_2(A)$ for three values of the number of the elements N. Compared to Table 9.1, we observe that the condition number of the matrix (10.1.9) grows with respect to its size much slower than the matrix in Example 9.1.1.

N	$\mathrm{Cond}_2(A)$
10^1	3.99E+01
10^2	4.05E+03
10^5	4.05E+09

TABLE 10.1. Condition numbers for the matrix (10.1.9)

10.1.2 High order elements and the condensation technique

We still consider the finite element method for solving the boundary value problem (10.1.1). This time we use piecewise quadratic functions. So the finite element space is

$$V_h = \{v_h \in V \mid v_h|_{K_i} \text{ is quadratic}\}.$$

Equivalently,

$$V_h = \{v_h \in C(\overline{\Omega}) \mid v_h|_{K_i} \text{ is quadratic, } v_h(0) = 0\}.$$

Let us introduce a basis for the space V_h. We denote the mid-points of the subintervals by $x_{i-1/2} = (x_{i-1} + x_i)/2$, $1 \le i \le N$. Associated with each node x_i, $1 \le i \le N-1$, we define

$$\phi_i(x) = \begin{cases} 2\,(x - x_{i-1})\,(x - x_{i-1/2})/h_i^2, & x \in [x_{i-1}, x_i], \\ 2\,(x_{i+1} - x)\,(x_{i+1/2} - x)/h_{i+1}^2, & x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise.} \end{cases} \tag{10.1.10}$$

Associated with x_N, we define

$$\phi_N(x) = \begin{cases} 2\,(x - x_{N-1})\,(x - x_{N-1/2})/h_N^2, & x \in [x_{N-1}, x_N], \\ 0, & \text{otherwise.} \end{cases} \tag{10.1.11}$$

We also need basis functions associated with the mid-points $x_{i-1/2}$, $1 \le i \le N$,

$$\psi_{i-1/2}(x) = \begin{cases} 4\,(x_i - x)\,(x - x_{i-1})/h_i^2, & x \in [x_{i-1}, x_i], \\ 0, & \text{otherwise.} \end{cases} \tag{10.1.12}$$

We notice that a mid-point basis function is non-zero only in one element. Now the finite element space can be represented as

$$V_h = \text{span}\,\{\phi_i, \psi_{i-1/2} \mid 1 \le i \le N\},$$

and we write

$$u_h = \sum_{j=1}^{N} u_j \phi_j + \sum_{j=1}^{N} u_{j-1/2} \psi_{j-1/2}.$$

The finite element system

$$\begin{cases} a(u_h, \phi_i) = \ell(\phi_i), \quad 1 \le i \le N, \\ a(u_h, \psi_{i-1/2}) = \ell(\psi_{i-1/2}), \quad 1 \le i \le N \end{cases}$$

can be written as, in the matrix/vector notation,

$$M_{11}\boldsymbol{u} + M_{12}\tilde{\boldsymbol{u}} = \boldsymbol{b}_1, \tag{10.1.13}$$

$$M_{21}\boldsymbol{u} + D_{22}\tilde{\boldsymbol{u}} = \boldsymbol{b}_2. \tag{10.1.14}$$

Here, $\boldsymbol{u} = (u_1, \ldots, u_N)^T$, $\tilde{\boldsymbol{u}} = (u_{1/2}, \ldots, u_{N-1/2})^T$, $M_{11} = (a(\phi_j, \phi_i))_{N\times N}$ is a tridiagonal matrix, $M_{12} = (a(\psi_{j-1/2}, \phi_i))_{N\times N}$ is a matrix with two diagonals, $M_{21} = M_{12}^T$, and $D_{22} = (a(\psi_{j-1/2}, \psi_{i-1/2}))_{N\times N}$ is a diagonal matrix with positive diagonal elements. We can eliminate $\tilde{\boldsymbol{u}}$ from the system (10.1.13)–(10.1.14) easily (both theoretically and practically). From (10.1.14), we have

$$\tilde{\boldsymbol{u}} = D_{22}^{-1}(\boldsymbol{b}_2 - M_{21}\boldsymbol{u}).$$

This relation is substituted into (10.1.13),

$$M\boldsymbol{u} = \boldsymbol{b}, \tag{10.1.15}$$

where $M = M_{11} - M_{12}D_{22}^{-1}M_{21}$ is a tridiagonal matrix, $\boldsymbol{b} = \boldsymbol{b}_1 - M_{12}D_{22}^{-1}\boldsymbol{b}_2$. It can be shown that M is positive definite.

As a result we see that for the finite element solution with quadratic elements, we only need to solve a tridiagonal system of order N, just like in the case of using linear elements in Subsection 10.1.1. The procedure of eliminating $\tilde{\boldsymbol{u}}$ from (10.1.13)–(10.1.14) to form a smaller size system (10.1.15) is called *condensation*. The key for the success of the condensation technique is that the supports of some basis functions are limited to a single element.

This condensation technique is especially useful in using high order elements to solve higher dimensional problems.

Another choice of the basis functions is $\{\phi_i\}_{1\le i\le N}$ defined in (10.1.3) and (10.1.4), together with $\{\psi_{i-1/2}\}_{1\le i\le N}$ defined in (10.1.12). A consequence of this choice is that the matrix M_{11} in (10.1.13) is the stiffness matrix of the finite element method associated with the linear elements. This basis is an example of hierarchical basis, i.e., when the local polynomial degree is increased, the basis functions for the low degree finite element space form a subset of the basis functions for the high degree finite element space. The use of hierarchical basis is beneficial in applying the finite element method when several finite element spaces are employed with increasing local polynomial degrees.

10.1.3 Reference element technique

Here we introduce the *reference element technique* which plays an important role for higher dimensional problems.

Consider a clamped beam, initially occupying the region $[0,1]$, which is subject to the action of a transversal force of density f. Denote u the deflection of the beam. Then the boundary value problem is

$$\begin{cases} u^{(4)} = f \quad \text{in } \Omega = (0,1), \\ u(0) = u'(0) = u(1) = u'(1) = 0. \end{cases} \tag{10.1.16}$$

The weak formulation of the problem is

$$u \in V, \quad \int_0^1 u''v''dx = \int_0^1 f\,v\,dx \quad \forall\, v \in V, \tag{10.1.17}$$

where $V = H_0^2(0,1)$. If we use the conforming finite element method, i.e., choose the finite element space V_h to be a subspace of V, then any function in V_h must be C^1 continuous. Suppose V_h consists of piecewise polynomials of degree less than or equal to p. The requirement that a finite element function be C^1 is equivalent to the C^1 continuity of the function across the interior nodal points $\{x_i\}_{i=1}^{N-1}$, which places $2\,(N-1)$ constraints. Additionally, the Dirichlet boundary conditions impose 4 constraints. Hence,

$$\dim\,(V_h) = (p+1)\,N - 2\,(N-1) - 4 = (p-1)\,N - 2.$$

Now it is evident that the polynomial degree p must be at least 2. However, with $p = 2$, we cannot construct basis functions with small supports. Thus we should choose p to be at least 3. For $p = 3$, our finite element space is taken to be

$$\begin{aligned} V_h = \{v_h \in C^1(\overline{\Omega}) \mid\ & v_h|_{K_i} \in \mathbb{P}_3(K_i),\ 1 \le i \le N, \\ & v_h(x) = v_h'(x) = 0 \text{ at } x = 0,1\}. \end{aligned}$$

It is then possible to construct basis functions with small supports using interpolation conditions of the function and its first derivative at the interior nodes $\{x_i\}_{i=1}^{N-1}$. More precisely, associated with each interior node x_i, there are two basis functions ϕ_i and ψ_i satisfying the interpolation conditions

$$\begin{aligned} &\phi_i(x_j) = \delta_{ij}, \quad \phi_i'(x_j) = 0, \\ &\psi_i(x_j) = 0, \quad \psi_i'(x_j) = \delta_{ij}. \end{aligned}$$

A more plausible approach to constructing the basis functions is to use the reference element technique. To this end, let us choose $\hat{K} = [0,1]$ as the reference element. Then the mapping

$$F_i : \hat{K} \to K_i,\ F_i(\hat{x}) = x_{i-1} + h_i\hat{x}$$

is a bijection between $\hat{K}$ and K_i. Over the reference element $\hat{K}$, we construct cubic functions Φ_0, Φ_1, Ψ_0 and Ψ_1 satisfying the interpolation con-

ditions

$$\begin{array}{llll}
\Phi_0(0)=1, & \Phi_0(1)=0, & \Phi_0'(0)=0, & \Phi_0'(1)=0,\\
\Phi_1(0)=0, & \Phi_1(1)=1, & \Phi_1'(0)=0, & \Phi_1'(1)=0,\\
\Psi_0(0)=0, & \Psi_0(1)=0, & \Psi_0'(0)=1, & \Psi_0'(1)=0,\\
\Psi_1(0)=0, & \Psi_1(1)=0, & \Psi_1'(0)=0, & \Psi_1'(1)=1.
\end{array}$$

It is not difficult to find these functions,

$$\begin{aligned}
\Phi_0(\hat{x}) &= (1+2\,\hat{x})\,(1-\hat{x})^2,\\
\Phi_1(\hat{x}) &= (3-2\,\hat{x})\,\hat{x}^2,\\
\Psi_0(\hat{x}) &= \hat{x}\,(1-\hat{x})^2,\\
\Psi_1(\hat{x}) &= -(1-\hat{x})\,\hat{x}^2.
\end{aligned}$$

These functions, defined on the reference element, are called *shape functions*. With the shape functions, it is an easy matter to construct the basis functions with the aid of the mapping functions $\{F_i\}_{i=1}^{N-1}$. We have

$$\phi_i(x)=\begin{cases}
\Phi_1(F_i^{-1}(x)), & x\in K_i,\\
\Phi_0(F_{i+1}^{-1}(x)), & x\in K_{i+1},\\
0, & \text{otherwise},
\end{cases}$$

and

$$\psi_i(x)=\begin{cases}
h_i\Psi_1(F_i^{-1}(x)), & x\in K_i,\\
h_{i+1}\Psi_0(F_{i+1}^{-1}(x)), & x\in K_{i+1},\\
0, & \text{otherwise}.
\end{cases}$$

Once the basis functions are available, it is a routine work to form the finite element system. We emphasize that the computations of the stiffness matrix and the load are done on the reference element. For example, by definition,

$$a_{i-1,i}=\int_0^1(\phi_{i-1})''(\phi_i)''dx=\int_{K_i}(\phi_{i-1})''(\phi_i)''dx;$$

using the mapping function F_i and the definition of the basis functions, we have

$$\begin{aligned}
a_{i-1,i} &= \int_{\hat{K}}(\Phi_0)''h_i^{-2}(\Phi_1)''h_i^{-2}\,h_i d\hat{x}\\
&= \frac{1}{h_i^3}\int_{\hat{K}}6\,(2\,\hat{x}-1)\,6\,(1-2\,\hat{x})\,d\hat{x}\\
&= -\frac{12}{h_i^3}.
\end{aligned}$$

For higher dimensional problems, the use of the reference element technique is essential for both theoretical error analysis and practical implementation of the finite element method. The computations of the stiffness matrix and the load vector involve a large number of integrals which usually cannot be computed analytically. With the reference element technique, all the integrals are computed on a single region—the reference element, and therefore numerical quadratures are needed on the reference element only. Also, we will see how the reference element technique is used to derive error estimates for the finite element interpolations.

Exercise 10.1.1 In Subsection 10.1.1, we computed the stiffness matrix for the case of a uniform partition. Find the stiffness matrix when the partition is non-uniform.

Exercise 10.1.2 Use the fact that the coefficient matrix of the system (10.1.13) and (10.1.14) is symmetric, positive definite to show that the coefficient matrix of the system (10.1.15) is symmetric, positive definite.

Exercise 10.1.3 Show that in solving (10.1.16) with a conforming finite element method with piecewise polynomials of degree less than or equal to 2, it is impossible to construct basis functions with a union of two neighboring elements as support.

10.2 Basics of the finite element method

We have seen from the one-dimensional examples in the preceding section that there are some typical steps in a finite element solution of a boundary value problem. First we need a weak formulation of the boundary value problem; this topic was discussed in Chapter 8. Then we need a partition (or triangulation) of the domain into sub-domains called elements. Associated with the partition, we define a finite element space. Further, we choose basis functions for the finite element space. The basis functions should have small supports so that the resulting stiffness matrix is sparse. With the basis functions defined, the finite element system can be formed. The reference element technique is used from time to time in this process.

In this section, we discuss basic aspects of finite elements. We restrict our discussion to two-dimensional problems; most of the discussion can be extended to higher-dimensional problems straightforwardly. We will assume the domain Ω is a polygon so that it can be partitioned into straight-sided triangles and quadrilaterals. When Ω is a general domain with a curved boundary, it cannot be partitioned into straight-sided triangles and quadrilaterals, and usually curved-sided elements need to be used. The reader is referred to [52, 53] for some detailed discussion on the use of the finite element method in this case. We will emphasize the use of the reference

element technique to estimate the error; for this reason, we need a particular structure on the finite elements, namely, we will consider only affine families of finite elements. In general, bilinear functions are needed for a bijective mapping between a four-sided reference element and a general quadrilateral. So a further restriction is that we will mostly use triangular elements, for any triangle is affine equivalent to a fixed triangle—the reference triangle.

First, we need a triangulation of the domain $\overline{\Omega}$ into subsets. Here we consider triangular subsets. We say $\mathcal{T}_h = \{K\}$ is a triangulation, a mesh, or a partition of the domain $\overline{\Omega}$ into triangular elements if the following properties hold:

1. $\overline{\Omega} = \cup_{K \in \mathcal{T}_h} K$.
2. Each K is a triangle.
3. For distinct $K_1, K_2 \in \mathcal{T}_h$, $\overset{\circ}{K}_1 \cap \overset{\circ}{K}_2 = \emptyset$.
4. For distinct $K_1, K_2 \in \mathcal{T}_h$, $K_1 \cap K_2$ is empty, or a common vertex, or a common side of K_1 and K_2.

The second property is introduced just for convenience in the following discussion; rectangular and general quadrilateral elements are also widely used. In the third property, $\overset{\circ}{K}$ denotes the interior of a set K. The fourth property is called a *regularity condition*. Each $K \in \mathcal{T}_h$ is called an element. For a triangulation of a three-dimensional domain into tetrahedral, hexahedral or pentahedral elements, the regularity condition requires that the intersection of two distinct elements is empty, or a common vertex, or a common side, or a common face of the two elements. For an arbitrary element K, we denote

$$h_K = \operatorname{diam}(K) = \max\{\|\boldsymbol{x} - \boldsymbol{y}\| \mid \boldsymbol{x}, \boldsymbol{y} \in K\}$$

and

$$\rho_K = \text{diameter of the largest sphere inscribed in } K.$$

Since K is a triangular element, h_K is the length of the longest side. The quantity h_K describes the size of K, while the ratio h_K/ρ_K is an indication whether the element is flat. We denote $h = \max_{K \in \mathcal{T}_h} h_K$ for the meshsize of the partition $\mathcal{T}_h$.

We will explain the ideas with the continuous linear elements, and then move on to a general framework.

10.2.1 Continuous linear elements

Suppose the boundary value problem over Ω under consideration is of second order. Then the Sobolev space $H^1(\Omega)$ or its subsets are used in the

weak formulation. As an example, we consider solving the following Neumann boundary value problem:

$$-\Delta u + u = f \quad \text{in } \Omega, \tag{10.2.1}$$

$$\frac{\partial u}{\partial \nu} = g \quad \text{on } \Gamma, \tag{10.2.2}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$ are given. The weak formulation of the boundary value problem is

$$u \in V, \quad \int_\Omega (\nabla u \cdot \nabla v + u\, v)\, dx = \int_\Omega f\, v\, dx + \int_\Gamma g\, v\, ds \quad \forall\, v \in V, \tag{10.2.3}$$

where the space $V = H^1(\Omega)$. We construct a linear element space V_h of continuous piecewise linear functions for V:

$$V_h = \{v_h \in V \mid v_h|_K \in \mathbb{P}_1 \ \forall\, K \in \mathcal{T}_h\}. \tag{10.2.4}$$

Since the restriction of v_h on each element K is smooth, a necessary and sufficient condition for $v_h \in H^1(\Omega)$ is $v_h \in C(\overline{\Omega})$ (see Examples 7.2.7 and 7.2.8 in Chapter 7). So equivalently,

$$V_h = \left\{v_h \in C(\overline{\Omega}) \mid v_h|_K \in \mathbb{P}_1 \ \forall\, K \in \mathcal{T}_h\right\}. \tag{10.2.5}$$

Let v_h be a piecewise linear function. Then $v_h \in C(\overline{\Omega})$ if and only if

$$v_h|_{K_1} = v_h|_{K_2} \text{ on } K_1 \cap K_2, \ \forall\, K_1, K_2 \in \mathcal{T}_h \text{ with } K_1 \cap K_2 \neq \emptyset. \tag{10.2.6}$$

For (10.2.6) to hold, it is natural to define $v_h|_K$ by its values at the three vertices of K, for any $K \in \mathcal{T}_h$. Thus, let us determine any function $v_h \in V_h$ by its values at all the vertices. Let $\{\boldsymbol{x}_i\}_{i=1}^{N_h} \subset \overline{\Omega}$ be the set of all the vertices of the elements $K \in \mathcal{T}_h$. Then a basis of the space V_h is $\{\phi_i\}_{1 \le i \le N_h}$ where the basis function ϕ_i is associated with the vertex $\boldsymbol{x}_i$, i.e., it satisfies the following conditions:

$$\phi_i \in V_h, \quad \phi_i(\boldsymbol{x}_j) = \delta_{ij}, \ 1 \le i, j \le N_h. \tag{10.2.7}$$

We can then write

$$V_h = \text{span}\{\phi_i\}_{1 \le i \le N_h}. \tag{10.2.8}$$

For any element $K \in \mathcal{T}_h$ containing $\boldsymbol{x}_i$ as a vertex, ϕ_i is a linear function on K, whereas if $\boldsymbol{x}_i$ is not a vertex of K, then $\phi_i(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in K$. If we write

$$v_h = \sum_{i=1}^{N_h} v_i \phi_i, \tag{10.2.9}$$

then using the property (10.2.7) we obtain

$$v_i = v_h(\boldsymbol{x}_i), \quad 1 \le i \le N_h. \tag{10.2.10}$$

We say the vertices $\{\boldsymbol{x}_i\}_{1\le i\le N_h}$ are the *nodes* of the linear element space V_h defined in (10.2.4) or (10.2.5). The finite element method for (10.2.3) is

$$u_h \in V_h, \quad \int_\Omega (\nabla u_h \cdot \nabla v_h + u_h v_h)\, dx = \int_\Omega f\, v_h dx + \int_\Gamma g\, v_h ds \quad \forall\, v_h \in V_h. \tag{10.2.11}$$

Write the finite element solution as

$$u_h = \sum_{j=1}^{N_h} u_j \phi_j.$$

Then from (10.2.11), we have the following discrete system for the coefficients $\{u_j\}_{1\le j\le N_h}$:

$$\sum_{j=1}^{N_h} u_j \int_\Omega (\nabla \phi_j \cdot \nabla \phi_i + \phi_j \phi_i)\, dx = \int_\Omega f\, \phi_i dx + \int_\Gamma g\, \phi_i ds, \quad 1 \le i \le N_h. \tag{10.2.12}$$

We now discuss construction of the basis functions $\{\phi_i\}_{1\le i\le N_h}$. For this purpose, it is convenient to use *barycentric coordinates*. Let K be a triangle and denote its three vertices by $\boldsymbol{a}_j = (a_{1j}, a_{2j})^T$, $1 \le j \le 3$, numbered counter-clockwise. We define the barycentric coordinates $\{\lambda_j(\boldsymbol{x})\}_{1\le j\le 3}$, associated with the vertices $\{\boldsymbol{a}_j\}_{1\le j\le 3}$, to be the affine functions satisfying

$$\lambda_j(\boldsymbol{a}_i) = \delta_{ij}, \quad 1 \le i, j \le 3. \tag{10.2.13}$$

Then by the uniqueness of linear interpolation, we have

$$\sum_{i=1}^{3} \lambda_i(\boldsymbol{x})\, \boldsymbol{a}_i = \boldsymbol{x}, \tag{10.2.14}$$

$$\sum_{i=1}^{3} \lambda_i(\boldsymbol{x}) = 1. \tag{10.2.15}$$

The equations (10.2.14) and (10.2.15) constitute a linear system for the three unknowns $\{\lambda_i\}_{i=1}^3$,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix}. \tag{10.2.16}$$

Note that the determinant of the coefficient matrix A of this system is twice the area of the triangle K. Since the triangle K is non-degenerate (i.e., the interior of K is nonempty), the matrix A is non-singular and so the system (10.2.16) uniquely determines the barycentric coordinates $\{\lambda_i\}_{i=1}^3$.

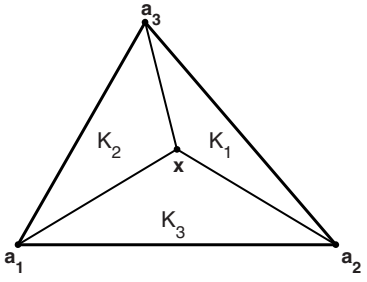

FIGURE 10.2. Interpretation of barycentric coordinates

By Cramer's rule, we have the following formulas:

$$\lambda_1(\boldsymbol{x}) = \frac{1}{\det A}\begin{vmatrix} x_1 & a_{12} & a_{13} \\ x_2 & a_{22} & a_{23} \\ 1 & 1 & 1 \end{vmatrix},$$

$$\lambda_2(\boldsymbol{x}) = \frac{1}{\det A}\begin{vmatrix} a_{11} & x_1 & a_{13} \\ a_{21} & x_2 & a_{23} \\ 1 & 1 & 1 \end{vmatrix},$$

$$\lambda_3(\boldsymbol{x}) = \frac{1}{\det A}\begin{vmatrix} a_{11} & a_{12} & x_1 \\ a_{21} & a_{22} & x_2 \\ 1 & 1 & 1 \end{vmatrix}.$$

For $\boldsymbol{x} \in K$, let K_1 be the triangle with the vertices $\boldsymbol{x}$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$, and similarly define two other triangles K_2 and K_3 as shown in Figure 10.2. Then from the above formulas, we get the following geometric interpretation of the barycentric coordinates:

$$\lambda_i(\boldsymbol{x}) = \frac{\text{area}(K_i)}{\text{area}(K)}, \quad 1 \le i \le 3, \quad \boldsymbol{x} \in K.$$

Since $\{\lambda_i\}_{i=1}^3$ are affine functions and $\lambda_i(\boldsymbol{a}_j) = \delta_{ij}$, the following relation holds:

$$\lambda_i(t_1\boldsymbol{a}_1 + t_2\boldsymbol{a}_2 + (1-t_1-t_2)\boldsymbol{a}_3) = t_1\delta_{i1} + t_2\delta_{i2} + (1-t_1-t_2)\,\delta_{i3},\ 1 \le i \le 3.$$

This formula is useful in computing the barycentric coordinates of points on the plane. In particular, λ_1 is a constant along any line parallel to the side $\overline{\boldsymbol{a}_2\boldsymbol{a}_3}$, and on such a parallel line, the value of λ_1 is proportional to the signed scaled distance of the line containing $\overline{\boldsymbol{a}_2\boldsymbol{a}_3}$ (the sign is positive for points on the same side as $\boldsymbol{a}_1$ of the line containing $\overline{\boldsymbol{a}_2\boldsymbol{a}_3}$, and negative on the other side). See Figure 10.3, where $\boldsymbol{a}_{12} = (\boldsymbol{a}_1 + \boldsymbol{a}_2)/2$ and $\boldsymbol{a}_{31} = (\boldsymbol{a}_3 + \boldsymbol{a}_1)/2$ are side mid-points.

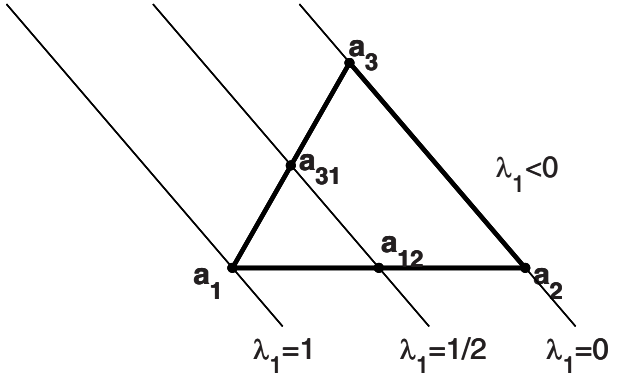

FIGURE 10.3. Value of barycentric coordinate λ_1

A basis function ϕ_i is non-zero on an element $K \in \mathcal{T}_h$ if and only if the associated node $\boldsymbol{x}_i$ is a vertex of K. Without loss of generality, let $\boldsymbol{a}_1$ be the node associated with ϕ_i. Then on K, $\phi_i(\boldsymbol{x}) = \lambda_1(\boldsymbol{x})$. Thus, with the barycentric coordinates, it is easy to compute the basis functions $\{\phi_i\}_{1\le i\le N_h}$ defined in (10.2.7). Also, note that over the element K, the function v_h of (10.2.9) is given by the simple formula

$$v_h = \sum_{i=1}^{3} v_h(\boldsymbol{a}_i)\,\lambda_i. \tag{10.2.17}$$

Back to the finite element system (10.2.12), we notice that for general functions f and g, the right hand sides can be computed only through numerical integrations. As an example, consider the calculation of the intergral $\int_\Omega f\,\phi_i dx$, which is the summation of elemental level intergrals of the form $\int_K f\,\phi_i dx$ where ϕ_i is non-zero on K. Rather than doing numerical integration on each element, we introduce a change of variables so that all the elemental level integrals are calculated over a fixed triangle $\hat{K}$, called a reference element. In this way, we only need numerical quadratures over $\hat{K}$ in computing integrals of the form $\int_\Omega f\,\phi_i dx$. The reference element for triangular elements is usually taken to be either an equilateral or right isosceles triangle, as shown in Figure 10.4.

For the equilateral triangular reference element with $\hat{\boldsymbol{a}}_1 = (-1, 0)$, $\hat{\boldsymbol{a}}_2 = (1, 0)$, and $\hat{\boldsymbol{a}}_3 = (0, \sqrt{3})$, the barycentric coordinates are

$$\begin{aligned}
\hat{\lambda}_1(\hat{\boldsymbol{x}}) &= \frac{1}{2}\left(1 - \hat{x}_1 - \frac{\hat{x}_2}{\sqrt{3}}\right),\\
\hat{\lambda}_2(\hat{\boldsymbol{x}}) &= \frac{1}{2}\left(1 + \hat{x}_1 - \frac{\hat{x}_2}{\sqrt{3}}\right),\\
\hat{\lambda}_3(\hat{\boldsymbol{x}}) &= \frac{\hat{x}_2}{\sqrt{3}}.
\end{aligned}$$

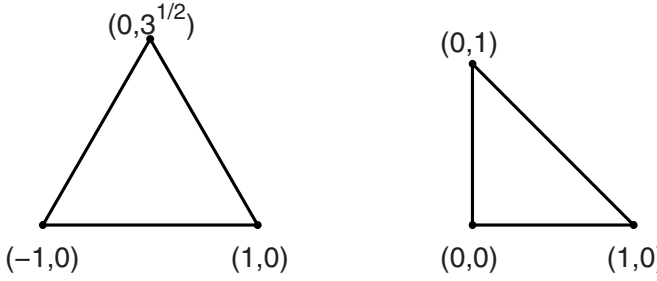

FIGURE 10.4. Reference triangular elements in $\mathbb{R}^2$

For the right isosceles triangular reference element with $\hat{\boldsymbol{a}}_1 = (0,0)$, $\hat{\boldsymbol{a}}_2 = (1,0)$, and $\hat{\boldsymbol{a}}_3 = (0,1)$, the barycentric coordinates are

$$\hat{\lambda}_1(\hat{\boldsymbol{x}}) = 1 - \hat{x}_1 - \hat{x}_2, \quad \hat{\lambda}_2(\hat{\boldsymbol{x}}) = \hat{x}_1, \quad \hat{\lambda}_3(\hat{\boldsymbol{x}}) = \hat{x}_2.$$

For definiteness, let us choose the right isosceles triangle reference element in the following discussion. We construct an affine mapping F_K from $\hat{K}$ to K such that $\boldsymbol{a}_i = F_K(\hat{\boldsymbol{a}}_i)$, $1 \le i \le 3$. It is easy to verify that

$$\boldsymbol{x} = F_K(\hat{\boldsymbol{x}}) = \boldsymbol{a}_1 + B_K \hat{\boldsymbol{x}}, \tag{10.2.18}$$

where the matrix

$$B_K = \begin{pmatrix} a_{12} - a_{11} & a_{13} - a_{11} \\ a_{22} - a_{21} & a_{23} - a_{21} \end{pmatrix}. \tag{10.2.19}$$

From now on, we will relate $\boldsymbol{x} \in K$ and $\hat{\boldsymbol{x}} \in \hat{K}$ by the relation (10.2.18). Moreover, we will relate a function v defined on K with a function $\hat{v}$ defined on $\hat{K}$ by the relation

$$v(\boldsymbol{x}) = \hat{v}(\hat{\boldsymbol{x}}). \tag{10.2.20}$$

An integral over an element K can be transformed to one over the reference element $\hat{K}$ as follows:

$$\int_K v(\boldsymbol{x})\, dx = \det(B_K) \int_{\hat{K}} v(F_K(\hat{\boldsymbol{x}}))\, d\hat{x}, \tag{10.2.21}$$

which is then approximated by applying a numerical quadrature over the fixed region $\hat{K}$. The boundary integral term in (10.2.12) can be handled similarly, and the calculations are done on the sides of the reference element $\hat{K}$.

Actually, the finite element space (10.2.8) can be constructed from a single function space over $\hat{K}$ together with the mappings (10.2.18). We use the symbol $\hat{X}$ for the set of functions $\hat{v} \in \mathbb{P}_1(\hat{K})$ that are determined by their values at the three vertices $\hat{\boldsymbol{a}}_1$, $\hat{\boldsymbol{a}}_2$, and $\hat{\boldsymbol{a}}_3$. Introduce the basis

functions $\hat{\phi}_i(\hat{\boldsymbol{x}}) = \hat{\lambda}_i(\hat{\boldsymbol{x}})$, $1 \le i \le 3$, associated with the vertices. Any $\hat{v} \in \hat{X}$ can be expressed as

$$\hat{v}(\hat{\boldsymbol{x}}) = \sum_{i=1}^{3} \hat{v}(\hat{\boldsymbol{a}}_i)\, \hat{\phi}_i(\hat{\boldsymbol{x}}).$$

Now consider a general element K with the three vertices $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$. Then for any $v_h \in V_h$, $v_h|_K \in \mathbb{P}_1(K)$ can be written as

$$v_h(\boldsymbol{x}) = \sum_{i=1}^{3} v_h(\boldsymbol{a}_i)\, \lambda_i(\boldsymbol{x}) = \sum_{i=1}^{3} v_h(\boldsymbol{a}_i)\, \hat{\phi}_i(\hat{\boldsymbol{x}}), \quad \boldsymbol{x} \in K.$$

In this way, the restriction of any function $v_h \in V_h$ on any element K can be obtained from a corresponding function in $\hat{X}$. Since numerical integrations in constructing finite element systems are done on the reference element, the finite element method is completely determined by the space $\hat{X}$ over the reference element $\hat{K}$ and the mapping functions $\{F_K\}_{K \in \mathcal{T}_h}$. Consequently, it is actually not necessary to construct the basis functions $\{\phi_i\}_{1 \le i \le N_h}$ of the finite element space V_h.

Consider a mixed boundary value problem where we have a homogeneous Dirichlet boundary condition on part of the boundary Γ_1 (relatively closed) and a Neumann boundary condition on the other part of the boundary Γ_2 (relatively open), then the triangulation should be compatible with the splitting $\partial\Omega = \Gamma_1 \cup \Gamma_2$, i.e., if an element K has one side on the boundary, then that side must belong entirely to either Γ_1 or $\overline{\Gamma_2}$. The corresponding linear element space is then constructed from the basis functions associated with the interior nodes together with those boundary nodes located on Γ_1. In other words, if $\{\boldsymbol{x}_i\}_{1 \le i \le N_{h,0}}$, $N_{h,0} < N_h$, are all the nodes in Ω and on Γ_2, then the linear element space to be used is

$$V_h = \operatorname{span}\{\phi_i\}_{1 \le i \le N_{h,0}} \tag{10.2.22}$$

instead of (10.2.8).

10.2.2 Affine-equivalent finite elements

Given finite element partitions of the polygon $\overline{\Omega}$, we now consider construction of finite element spaces based on affine-equivalent finite elements. We introduce a function space $\hat{X}$ over the reference element $\hat{K}$, including a description on how a function in $\hat{X}$ is determined, and then construct a corresponding function space on a general element K by using the mapping function (10.2.18). Although it is possible to choose any finite dimensional function space as $\hat{X}$, the overwhelming choice for $\hat{X}$ for practical use is a polynomial space.

It is convenient to use the barycentric coordinates to represent polynomials. For example, if $\hat{X}_1$ consists of functions from $\mathbb{P}_1(\hat{K})$ that are determined

by their values at the vertices $\{\hat{\boldsymbol{a}}_i\}_{i=1}^3$. Then for any $\hat{v} \in \hat{X}_1$, we have the representation

$$\hat{v}(\hat{\boldsymbol{x}}) = \sum_{i=1}^{3} \hat{v}(\hat{\boldsymbol{a}}_i)\, \hat{\lambda}_i(\hat{\boldsymbol{x}}).$$

In this case, the vertices $\{\hat{\boldsymbol{a}}_i\}_{i=1}^3$ are called the nodes, the function values $\{\hat{v}(\hat{\boldsymbol{a}}_i)\}_{i=1}^3$ are called the parameters (used to determine the linear function).

A quadratic function has six coefficients and we need six interpolation conditions to determine it. For this, we introduce the side mid-points,

$$\hat{\boldsymbol{a}}_{ij} = \frac{1}{2}\,(\hat{\boldsymbol{a}}_i + \hat{\boldsymbol{a}}_j), \quad 1 \le i < j \le 3.$$

Then we introduce the space $\hat{X}_2$ of $\mathbb{P}_2(\hat{K})$ functions that are determined by their values at the vertices $\{\hat{\boldsymbol{a}}_i\}_{i=1}^3$ and the side mid-points $\{\hat{\boldsymbol{a}}_{ij}\}_{1\le i<j\le 3}$. For any $\hat{v} \in \hat{X}_2$, we have the representation formula

$$\hat{v}(\hat{\boldsymbol{x}}) = \sum_{i=1}^{3} \hat{v}(\hat{\boldsymbol{a}}_i)\, \hat{\lambda}_i(\hat{\boldsymbol{x}})\,(2\,\hat{\lambda}_i(\hat{\boldsymbol{x}}) - 1) + \sum_{1\le i<j\le 3} 4\,\hat{v}(\hat{\boldsymbol{a}}_{ij})\, \hat{\lambda}_i(\hat{\boldsymbol{x}})\, \hat{\lambda}_j(\hat{\boldsymbol{x}}). \tag{10.2.23}$$

This formula is derived from the observations that
(1) for each i, $1 \le i \le 3$, $\hat{\lambda}_i(\hat{\boldsymbol{x}})\,(2\,\hat{\lambda}_i(\hat{\boldsymbol{x}}) - 1)$ is a quadratic function that takes on the value 1 at $\hat{\boldsymbol{a}}_i$, and the value 0 at the other vertices and the side mid-points;
(2) for $1 \le i < j \le 3$, $4\,\hat{\lambda}_i(\hat{\boldsymbol{x}})\, \hat{\lambda}_j(\hat{\boldsymbol{x}})$ is a quadratic function that takes on the value 1 at $\hat{\boldsymbol{a}}_{ij}$, and the value 0 at the other side mid-points and the vertices. In this case, the vertices and the side mid-points are called the nodes, the function values at the nodes are called the parameters (used to determine the quadratic function).

For a positive integer k, let $\mathbb{P}_k(\hat{K})$ be the space of polynomials of degree less than or equal to k on $\hat{K}$. Note that the dimension of the space is $\dim \mathbb{P}_k(\hat{K}) = (k+2)(k+1)/2$. To uniquely determine a $\hat{v} \in \mathbb{P}_k(\hat{K})$, we can use the values of v at the following nodal points:

$$\hat{N}_k = \left\{ \sum_{i=1}^{3} t_i \hat{\boldsymbol{a}}_i \;\middle|\; \sum_{i=1}^{3} t_i = 1,\ t_i \in \left\{\frac{j}{k}\right\}_{0\le j\le k},\ 1 \le i \le 3 \right\}. \tag{10.2.24}$$

Note that this set contains $(k+2)(k+1)/2$ points. The following result is proved in [179].

Proposition 10.2.1 *A polynomial $\hat{v} \in \mathbb{P}_k(\hat{K})$ is uniquely determined by its values at the points in the set $\hat{N}_k$.*

We then introduce the space $\hat{X}_k$ consisting of $\mathbb{P}_k(\hat{K})$ functions determined by their values at the points of the set $\hat{N}_k$. The cases $k = 1$ and $k = 2$ have been introduced previously.

Above, the parameters are function values at the nodes. The corresponding finite elements to be constructed later are called Lagrange finite elements. Lagrange finite elements are the natural choice in solving second-order boundary value problems, where weak formulations only need the use of weak derivatives of the first order, and hence only the continuity of finite element functions across interelement boundaries is required. It is also possible to use other types of parameters to determine polynomials. For example, we may choose some parameters to be derivatives of the function at some nodes; in this case, we will get Hermite finite elements. We can construct Hermite finite elements which are globally continuously differentiable; such Hermite finite elements can be used to solve fourth-order boundary value problems, as was discussed in Subsection 10.1.3 in the context of a one-dimensional problem. For some applications, it may be advantageous to use average values of the function along the sides. Different selections of the parameters lead to different basis functions, and thus lead to different finite element system. Here we will focus on the discussion of Lagrange finite elements. The discussion can be extended to the case of higher degree polynomials and to domains of any (finite) dimension. A reader interested in a more complete discussion of general finite elements can consult the references [52, 53].

In general, let $\hat{X}$ be a finite dimensional space over $\hat{K}$, with a dimension $\dim \hat{X} = I$, such that any function $\hat{v} \in \hat{X}$ is uniquely determined by its values at the I nodes $\hat{\boldsymbol{x}}_1, \cdots, \hat{\boldsymbol{x}}_I \in \hat{K}$. Then we have the following formula

$$\hat{v}(\hat{\boldsymbol{x}}) = \sum_{i=1}^{I} \hat{v}(\hat{\boldsymbol{x}}_i)\, \hat{\phi}_i(\hat{\boldsymbol{x}}).$$

The functions $\{\hat{\phi}_i\}_{i=1}^{I}$ form a basis for the space $\hat{X}$ with the property

$$\hat{\phi}_i(\hat{\boldsymbol{x}}_j) = \delta_{ij}.$$

We will then define function spaces over a general element. For the *affine-equivalent* families of finite elements, there exists one or several reference elements, $\hat{K}$, such that each element K is the image of $\hat{K}$ under an invertible affine mapping $F_K : \hat{K} \to K$ of the form

$$F_K(\hat{\boldsymbol{x}}) = \boldsymbol{T}_K \hat{\boldsymbol{x}} + \boldsymbol{b}_K. \tag{10.2.25}$$

The mapping F_K is a bijection between $\hat{K}$ and K, $\boldsymbol{T}_K$ is an invertible 2×2 matrix and $\boldsymbol{b}_K$ is a translation vector.

For each element K, we establish a correspondence between functions defined on K and $\hat{K}$ through the use of the affine mapping F_K. For any

function v defined on K, $\hat{v}$ denotes the corresponding function defined on $\hat{K}$ through $\hat{v} = v \circ F_K$. Conversely, for any function $\hat{v}$ on $\hat{K}$, we let v be the function on K defined by $v = \hat{v} \circ F_K^{-1}$. Thus we have the relation

$$v(\boldsymbol{x}) = \hat{v}(\hat{\boldsymbol{x}}) \quad \forall\, \boldsymbol{x} \in K,\ \hat{\boldsymbol{x}} \in \hat{K},\ \text{with } \boldsymbol{x} = F_K(\hat{\boldsymbol{x}}).$$

We then define a finite dimensional function space X_K formally by the formula

$$X_K = \hat{X} \circ F_K^{-1} \tag{10.2.26}$$

where any function $v \in X_K$ corresponds to a function $\hat{v} \in \hat{X}$ with $v = \hat{v} \circ F_K^{-1}$. Implicit in the definition (10.2.26) is that functions in X_K are determined in the same way as functions in $\hat{X}$; e.g., if $\hat{X}$ consists of $\mathbb{P}_1(\hat{K})$ functions determined by their values at the three vertices of $\hat{K}$, then X_K consists of $\mathbb{P}_1(K)$ functions determined by their values at the three vertices of K. An immediate consequence of this definition is that if $\hat{X}$ contains polynomials of certain degree, then X_K contains polynomials of the same degree.

Using the nodal points $\hat{\boldsymbol{x}}_i$, $1 \le i \le I$, of $\hat{K}$, we introduce the nodal points $\boldsymbol{x}_i^K$, $1 \le i \le I$, of K defined by

$$\boldsymbol{x}_i^K = F_K(\hat{\boldsymbol{x}}_i), \quad i = 1, \ldots, I. \tag{10.2.27}$$

Recall that $\{\hat{\phi}_i\}_{i=1}^I$ are the basis functions of the space $\hat{X}$ associated with the nodal points $\{\hat{\boldsymbol{x}}_i\}_{i=1}^I$ with the property that

$$\hat{\phi}_i(\hat{\boldsymbol{x}}_j) = \delta_{ij}.$$

We define

$$\phi_i^K = \hat{\phi}_i \circ F_K^{-1}, \quad i = 1, \ldots, I.$$

Then the functions $\{\phi_i^K\}_{i=1}^I$ have the property that

$$\phi_i^K(\boldsymbol{x}_j^K) = \delta_{ij}.$$

Hence, $\{\phi_i^K\}_{i=1}^I$ form a set of local basis functions on K.

We now present a result on the affine transformation (10.2.25), which will be used in estimating finite element interpolation errors. The matrix norm is the spectral norm, i.e., the operator matrix norm induced by the Euclidean vector norm.

Lemma 10.2.2 *For the affine map $F_K : \hat{K} \to K$ defined by* (10.2.25)*, we have the bounds*

$$\|\boldsymbol{T}_K\| \le \frac{h_K}{\hat{\rho}} \quad \textit{and} \quad \|\boldsymbol{T}_K^{-1}\| \le \frac{\hat{h}}{\rho_K}.$$

Proof. By definition of the matrix norm,

$$\|\boldsymbol{T}_K\| = \sup\left\{ \frac{\|\boldsymbol{T}_K\hat{\boldsymbol{x}}\|}{\|\hat{\boldsymbol{x}}\|} \;\middle|\; \hat{\boldsymbol{x}} \neq \boldsymbol{0} \right\}.$$

Let us rewrite it in the equivalent form

$$\|\boldsymbol{T}_K\| = \hat{\rho}^{-1} \sup\{\|\boldsymbol{T}_K\hat{\boldsymbol{z}}\| \mid \|\hat{\boldsymbol{z}}\| = \hat{\rho}\}$$

by taking $\hat{\boldsymbol{z}} = \hat{\rho}\hat{\boldsymbol{x}}/\|\hat{\boldsymbol{x}}\|$. Now for any $\hat{\boldsymbol{z}}$ with $\|\hat{\boldsymbol{z}}\| = \hat{\rho}$, pick up any two vectors $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ that lie on the largest sphere $\hat{S}$ of diameter $\hat{\rho}$, which is inscribed in $\hat{K}$, such that $\hat{\boldsymbol{z}} = \hat{\boldsymbol{x}} - \hat{\boldsymbol{y}}$. Then

$$\begin{aligned}
\|\boldsymbol{T}_K\| &= \hat{\rho}^{-1} \sup\left\{ \|\boldsymbol{T}_K(\hat{\boldsymbol{x}} - \hat{\boldsymbol{y}})\| \mid \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \in \hat{S} \right\} \\
&= \hat{\rho}^{-1} \sup\left\{ \|(\boldsymbol{T}_K\hat{\boldsymbol{x}} + \boldsymbol{b}_K) - (\boldsymbol{T}_K\hat{\boldsymbol{y}} + \boldsymbol{b}_K)\| \mid \hat{\boldsymbol{x}}, \hat{\boldsymbol{y}} \in \hat{S} \right\} \\
&\leq \hat{\rho}^{-1} \sup\{\|\boldsymbol{x} - \boldsymbol{y}\| \mid \boldsymbol{x}, \boldsymbol{y} \in K\} \\
&\leq h_K/\hat{\rho}.
\end{aligned}$$

The second inequality follows from the first one by interchanging the roles played by K and $\hat{K}$. □

10.2.3 Finite element spaces

A global finite element function v_h is defined piecewise by the formula

$$v_h|_K \in X_K \quad \forall\, K \in \mathcal{T}_h.$$

We then define a finite element space corresponding to the triangulation $\mathcal{T}_h$:

$$X_h = \{v_h \mid v_h|_K \in X_K \;\forall\, K \in \mathcal{T}_h\}.$$

Note that the expression $v_h|_K \in X_K$ includes the way how $v_h|_K$ is determined. A natural question is whether $X_h \subset H^1(\Omega)$, in the context of solving a linear second-order elliptic boundary value problem. Thus we need to check if $v_h \in H^1(\Omega)$ holds for $v_h \in X_h$. Since the restriction of v_h on each element K is a smooth function, a necessary and sufficient condition for $v_h \in H^1(\Omega)$ is $v_h \in C(\overline{\Omega})$ (see Examples 7.2.7 and 7.2.8 in Chapter 7, also Exercise 10.2.4). Then $X_h \subset H^1(\Omega)$ if and only if $X_h \subset C(\overline{\Omega})$. We remark that the condition $v_h \in C(\overline{\Omega})$ is guaranteed if v_h is continuous across any interelement boundary, i.e., if the condition (10.2.6) is satisfied.

Assume $X_h \subset C(\overline{\Omega})$ is valid. Then for a second-order boundary value problem with Neumann boundary condition, $V = H^1(\Omega)$ and we use $V_h = X_h$ as the finite element space; for a second-order boundary value problem with the homogeneous Dirichlet condition, $V = H_0^1(\Omega)$ and we choose the finite element space to be

$$V_h = \{v_h \in X_h \mid v_h = 0 \text{ on } \Gamma\}.$$

We observe that if $\hat{X}$ consists of polynomials, then a function from the space X_h is a piecewise image of polynomials. In our special case of an affine family of finite elements, F_K is an affine mapping, and so any function in X_h is a piecewise polynomial. For more general mapping functions F_K (e.g., bilinear mapping functions used for quadrilateral elements), $v_h|_K$ is in general not a polynomial.

Example 10.2.3 For a general triangle $K \in \mathcal{T}_h$, let X_K consist of $\mathbb{P}_1(K)$ functions, determined by their values at the three vertices $\boldsymbol{a}_j$, $1 \le j \le 3$. Define the linear element space

$$X_h = \{v_h \mid v_h|_K \in X_K,\ K \in \mathcal{T}_h\}.$$

We use the barycentric coordinates $\lambda_j(\boldsymbol{x})$, $1 \le j \le 3$, corresponding to the three vertices $\boldsymbol{a}_j$, $1 \le j \le 3$. Note that

$$\lambda_i = \hat{\lambda}_i \circ F_K^{-1}, \quad 1 \le i \le 3.$$

On each element K, $v_h|_K$ is a linear function and we have

$$v_h|_K = \sum_{i=1}^{3} v_h(\boldsymbol{a}_i)\, \lambda_i, \quad K \in \mathcal{T}_h. \tag{10.2.28}$$

The local finite element function $v_h|_K$ can be obtained from a linear function defined on $\hat{K}$. For this, we let

$$\hat{v}(\hat{\boldsymbol{x}}) = \sum_{i=1}^{3} v_h(\boldsymbol{a}_i)\, \hat{\lambda}_i(\hat{\boldsymbol{x}}).$$

Then it is easily seen that $v_h|_K = \hat{v} \circ F_K^{-1}$.

The piecewise linear function v_h defined by (10.2.28) is globally continuous, or (10.2.6) is satisfied. To see this we only need to prove the continuity of v_h across $\gamma = K_1 \cap K_2$, the common side of two adjacent elements K_1 and K_2. Let us denote $v_1 = v_h|_{K_1}$ and $v_2 = v_h|_{K_2}$, and consider the difference $w = v_1 - v_2$. On γ, w is a linear function of one variable (the tangential variable along γ), and vanishes at the two end points of γ. Hence, $w = 0$, i.e. $v_1 = v_2$ along γ.

Summarizing, if we use the function values at the vertices to define the functions on each element K, then X_h is an affine-equivalent finite element space consisting of continuous piecewise linear functions. On each element, we have the representation formula (10.2.28). □

In general, for a positive integer k, consider a function v_h whose restriction on any element K belongs to $\mathbb{P}_k(K)$. Let F_K be the mapping between $\hat{K}$ and K. Then

$$\sum_{i=1}^{3} t_i \boldsymbol{a}_i = F_K\left(\sum_{i=1}^{3} t_i \hat{\boldsymbol{a}}_i\right) \quad \forall\, t_1, t_2, t_3 \in \mathbb{R} \text{ with } \sum_{i=1}^{3} t_i = 1.$$

So corresponding to (10.2.24), define a set of nodal points on K by

$$N_k(K) = \left\{ \sum_{i=1}^{3} t_i \boldsymbol{a}_i \;\Big|\; \sum_{i=1}^{3} t_i = 1,\ t_i \in \left\{ \frac{j}{k} \right\}_{0 \le j \le k},\ 1 \le i \le 3 \right\}. \quad (10.2.29)$$

Then if $v_h|_K$ is determined by its values on the set $N_k(K)$, the function $v_h \in C(\overline{\Omega})$. The verification of this claim for $k = 2$ is left as Exercise 10.2.4. The claim for a higher value k can be proved similarly.

Exercise 10.2.1 Over the reference triangle $\hat{K}$, use the set of nodes $\hat{N}_3$ as in (10.2.24) for cubic functions. Represent a cubic function in terms of its values at the points in $\hat{N}_3$.

Exercise 10.2.2 Discuss how to transform the integrals on the left hand side of the finite element equation in (10.2.12) to integrals over the reference element by using the barycentric coordinates $\{\hat{\lambda}_i\}_{1 \le i \le 3}$ and the mapping functions $\{F_K\}_{K \in \mathcal{T}_h}$.

Exercise 10.2.3 Assume $\overline{\Omega}$ is the union of some rectangles whose sides parallel to the coordinate axes. We partition $\overline{\Omega}$ into rectangular elements with sides parallel to the coordinate axes. In this case, the reference element is a square $\hat{K}$, say the unit square. The polynomial space over $\hat{K}$ is usually taken to be

$$\mathbb{Q}_{k,l}(\hat{K}) = \left\{ v(\hat{\boldsymbol{x}}) = \sum_{i \le k} \sum_{j \le l} a_{ij} \hat{x}_1^i \hat{x}_2^j :\ a_{ij} \in \mathbb{R},\ \hat{\boldsymbol{x}} \in \hat{K} \right\}$$

for non-negative integers k and l. For $k = l = 1$, we get the bilinear functions. Define a set of nodes and parameters over $\mathbb{Q}_{1,1}(\hat{K})$. Represent a bilinear function in terms of the parameters.

Exercise 10.2.4 Let $\mathcal{T}_h = \{K\}$ be a finite element partition of $\overline{\Omega}$ into triangles. Over the reference element $\hat{K}$, we determine $\hat{v} \in \mathbb{P}_2(\hat{K})$ by its values at the three vertices and the three side mid-points; see (10.2.23). Show that the corresponding finite element functions have the property $v_h \in C(\overline{\Omega})$.

Exercise 10.2.5 Let $\mathcal{T}_h = \{K\}$ be a finite element partition of $\overline{\Omega}$. Consider a function $\boldsymbol{v} : \Omega \to \mathbb{R}^d$. Assume $\boldsymbol{v}|_K \in H^1(K)^d$ for any $K \in \mathcal{T}_h$. Show that $\boldsymbol{v} \in H(\mathrm{div}; \Omega)$ (see (8.6.20) for the definition of this space) if and only if the following property holds: For any $K_1, K_2 \in \mathcal{T}_h$ with $\gamma = K_1 \cap K_2 \neq \emptyset$, denote $\boldsymbol{v}_1 = \boldsymbol{v}|_{K_1}$ and $\boldsymbol{v}_2 = \boldsymbol{v}|_{K_2}$, and let $\boldsymbol{\nu}_1$ and $\boldsymbol{\nu}_2$ be the unit outward normals on γ with respect to K_1 and K_2, then $\boldsymbol{v}_1 \cdot \boldsymbol{\nu}_1 + \boldsymbol{v}_2 \cdot \boldsymbol{\nu}_2 = 0$ on γ.

10.3 Error estimates of finite element interpolations

In this section, we present some estimates for the finite element interpolation error, which will be used in the next section to bound the error of the

finite element solution for a linear elliptic boundary value problem, through the application of Céa's inequality. The interpolation error estimates are derived through the use of the reference element technique. In other words, error estimates are first derived on the reference element, which are then translated to a general finite element. The results discussed in this section can be extended to the case of a general d-dimensional domain. Definitions of triangulation and finite elements in d-dimensional case are similar to those for the 2-dimensional case; see e.g. [52, 53].

10.3.1 Local interpolations

We first introduce an interpolation operator $\hat{\Pi}$ for continuous functions on $\hat{K}$. Recall that $\{\hat{\boldsymbol{x}}_i\}_{i=1}^I$ are the nodal points whereas $\{\hat{\phi}_i\}_{i=1}^I$ are the associated basis functions of the polynomial space $\hat{X}$ in the sense that $\hat{\phi}_i(\hat{\boldsymbol{x}}_j) = \delta_{ij}$ is valid. We define

$$\hat{\Pi} : C(\hat{K}) \to \hat{X}, \quad \hat{\Pi}\hat{v} = \sum_{i=1}^{I} \hat{v}(\hat{\boldsymbol{x}}_i)\hat{\phi}_i. \tag{10.3.1}$$

Evidently, $\hat{\Pi}\hat{v} \in \hat{X}$ is uniquely determined by the interpolation conditions

$$\hat{\Pi}\hat{v}(\hat{\boldsymbol{x}}_i) = \hat{v}(\hat{\boldsymbol{x}}_i), \quad i = 1, \ldots, I.$$

On any element K, we define similarly the interpolation operator Π_K by

$$\Pi_K : C(K) \to X_K, \quad \Pi_K v = \sum_{i=1}^{I} v(\boldsymbol{x}_i^K)\phi_i^K. \tag{10.3.2}$$

We see that $\Pi_K v \in X_K$ is uniquely determined by the interpolation conditions

$$\Pi_K v(\boldsymbol{x}_i^K) = v(\boldsymbol{x}_i^K), \quad i = 1, \ldots, I.$$

The following result explores the relation between the two interpolation operators. The result is of fundamental importance in error analysis of finite element interpolation.

Theorem 10.3.1 *For the two interpolation operators $\hat{\Pi}$ and Π_K introduced above, we have $\hat{\Pi}(\hat{v}) = (\Pi_K v) \circ F_K$, i.e., $\hat{\Pi}\hat{v} = \widehat{\Pi_K v}$.*

Proof. From the definition (10.3.2), we have

$$\Pi_K v = \sum_{i=1}^{I} v(\boldsymbol{x}_i^K)\phi_i^K = \sum_{i=1}^{I} \hat{v}(\hat{\boldsymbol{x}}_i)\phi_i^K.$$

Since $\phi_i^K \circ F_K = \hat{\phi}_i$, we obtain

$$(\Pi_K v) \circ F_K = \sum_{i=1}^{I} \hat{v}(\hat{\boldsymbol{x}}_i)\hat{\phi}_i = \hat{\Pi}\hat{v}. \qquad \square$$

Example 10.3.2 For a generic triangular element K, again use $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ to denote its vertices. Then with linear elements, for a continuous function v defined on K, its linear interpolant is

$$\Pi_K v(\boldsymbol{x}) = \sum_{i=1}^{3} v(\boldsymbol{a}_i)\, \lambda_i(\boldsymbol{x}), \quad \boldsymbol{x} \in K.$$

For the function $\hat{v} = v \circ F_K$ defined on the reference element $\hat{K}$, its linear interpolant is

$$\hat{\Pi}\hat{v}(\hat{\boldsymbol{x}}) = \sum_{i=1}^{3} \hat{v}(\hat{\boldsymbol{a}}_i)\, \hat{\lambda}_i(\hat{\boldsymbol{x}}), \quad \hat{\boldsymbol{x}} \in \hat{K}.$$

Here, $\hat{\boldsymbol{a}}_i = F_K^{-1}(\boldsymbol{a}_i)$, $1 \le i \le 3$, are the vertices of $\hat{K}$. Since $v(\boldsymbol{a}_i) = \hat{v}(\hat{\boldsymbol{a}}_i)$ by definition and $\lambda_i(\boldsymbol{x}) = \hat{\lambda}_i(\hat{\boldsymbol{x}})$ for $\boldsymbol{x} = F_K(\hat{\boldsymbol{x}})$, obviously the relation $\hat{\Pi}\hat{v} = \widehat{\Pi_K v}$ holds.

By using the mid-points of the sides, we can give a similar discussion of quadratic elements. □

10.3.2 Interpolation error estimates on the reference element

We first derive interpolation error estimates over the reference element.

Theorem 10.3.3 *Let k and m be nonnegative integers with $k > 0$, $k+1 \ge m$, and $\mathbb{P}_k(\hat{K}) \subset \hat{X}$. Let $\hat{\Pi}$ be the operators defined in* (10.3.1)*. Then there exists a constant c such that*

$$|\hat{v} - \hat{\Pi}\hat{v}|_{m,\hat{K}} \le c\, |\hat{v}|_{k+1,\hat{K}} \quad \forall\, \hat{v} \in H^{k+1}(\hat{K}). \tag{10.3.3}$$

Proof. Notice that $k > 0$ implies $H^{k+1}(\hat{K}) \hookrightarrow C(\hat{K})$, so $\hat{v} \in H^{k+1}(\hat{K})$ is continuous and $\hat{\Pi}\hat{v}$ is well-defined. From

$$\|\hat{\Pi}\hat{v}\|_{m,\hat{K}} \le \sum_{i=1}^{I} |\hat{v}(\hat{\boldsymbol{x}}_i)|\, \|\hat{\phi}_i\|_{m,\hat{K}} \le c\, \|\hat{v}\|_{C(\hat{K})} \le c\, \|\hat{v}\|_{k+1,\hat{K}},$$

we see that $\hat{\Pi}$ is a bounded operator from $H^{k+1}(\hat{K})$ to $H^m(\hat{K})$. By the assumption on the space $\hat{X}$, we have

$$\hat{\Pi}\hat{v} = \hat{v} \quad \forall\, \hat{v} \in \mathbb{P}_k(\hat{K}). \tag{10.3.4}$$

Using (10.3.4), we then have, for all $\hat{v} \in H^{k+1}(\hat{K})$ and all $\hat{p} \in \mathbb{P}_k(\hat{K})$,

$$\begin{aligned} |\hat{v} - \hat{\Pi}\hat{v}|_{m,\hat{K}} &\le \|\hat{v} - \hat{\Pi}\hat{v}\|_{m,\hat{K}} = \|\hat{v} - \hat{\Pi}\hat{v} + \hat{p} - \hat{\Pi}\hat{p}\|_{m,\hat{K}} \\ &\le \|(\hat{v} + \hat{p}) - \hat{\Pi}(\hat{v} + \hat{p})\|_{m,\hat{K}} \\ &\le \|\hat{v} + \hat{p}\|_{m,\hat{K}} + \|\hat{\Pi}(\hat{v} + \hat{p})\|_{m,\hat{K}} \\ &\le c\, \|\hat{v} + \hat{p}\|_{k+1,\hat{K}}. \end{aligned}$$

Since $\hat{p} \in \mathbb{P}_k(\hat{K})$ is arbitrary, we have

$$|\hat{v} - \hat{\Pi}\hat{v}|_{m,\hat{K}} \le c \inf_{\hat{p}\in\mathbb{P}_k(\hat{K})} \|\hat{v} + \hat{p}\|_{k+1,\hat{K}}.$$

By an application of Corollary 7.3.18, we get the estimate (10.3.3). □

In Theorem 10.3.3, the assumption $k > 0$ is made to warrant the continuity of an $H^{k+1}(\hat{K})$ function. In the d-dimensional case, this assumption is replaced by $k + 1 > d/2$. The property (10.3.4) is called a *polynomial invariance property* of the finite element interpolation operator.

10.3.3 Local interpolation error estimates

We now consider the finite element interpolation error over each element K. As in Theorem 10.3.3, we assume $k > 0$; this assumption ensures the property $H^{k+1}(K) \hookrightarrow C(K)$, and so for $v \in H^{k+1}(K)$, pointwise values $v(\boldsymbol{x})$ are meaningful. Let the projection operator $\Pi_K : H^{k+1}(K) \to X_K \subset H^m(K)$ be defined by (10.3.2).

To translate the result of Theorem 10.3.3 from the reference element $\hat{K}$ to the element K, we need to discuss the relations between Sobolev norms over the reference element and a general element.

Theorem 10.3.4 *Assume $\boldsymbol{x} = \boldsymbol{T}_K\hat{\boldsymbol{x}} + \boldsymbol{b}_K$ is a bijection from $\hat{K}$ to K. Then $v \in H^m(K)$ if and only if $\hat{v} \in H^m(\hat{K})$. Furthermore, for some constant c independent of K and $\hat{K}$, the estimates*

$$|\hat{v}|_{m,\hat{K}} \le c\,\|\boldsymbol{T}_K\|^m |\det \boldsymbol{T}_K|^{-1/2} |v|_{m,K} \tag{10.3.5}$$

and

$$|v|_{m,K} \le c\,\|\boldsymbol{T}_K^{-1}\|^m |\det \boldsymbol{T}_K|^{1/2} |\hat{v}|_{m,\hat{K}} \tag{10.3.6}$$

hold.

Proof. We only need to prove the inequality (10.3.5); the inequality (10.3.6) follows from (10.3.5) by interchanging the roles played by $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. Recall the multi-index notation: for $\alpha = (\alpha_1, \alpha_2)$,

$$\partial^\alpha_{\hat{\boldsymbol{x}}} = \frac{\partial^{|\alpha|}}{\partial \hat{x}_1^{\alpha_1} \partial \hat{x}_2^{\alpha_2}}, \quad \partial^\alpha_{\boldsymbol{x}} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}.$$

By a change of variables, we have

$$\begin{aligned} |\hat{v}|^2_{m,\hat{K}} &= \sum_{|\alpha|=m} \int_{\hat{K}} [\partial^\alpha_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}})]^2 \, d\hat{x} \\ &= \sum_{|\alpha|=m} \int_K \left[\partial^\alpha_{\hat{\boldsymbol{x}}} \hat{v}(F_K^{-1}(\boldsymbol{x}))\right]^2 |\det \boldsymbol{T}_K|^{-1} dx. \end{aligned}$$

Since the mapping function is affine, for any multi-index α with $|\alpha| = m$, we have

$$\partial^{\alpha}_{\hat{\boldsymbol{x}}}\hat{v} = \sum_{|\beta|=m} c_{\alpha,\beta}(\boldsymbol{T}_K)\,\partial^{\beta}_{\boldsymbol{x}} v,$$

where each $c_{\alpha,\beta}(\boldsymbol{T}_K)$ is a product of m entries of the matrix $\boldsymbol{T}_K$. Thus

$$\sum_{|\alpha|=m} |\partial^{\alpha}_{\hat{\boldsymbol{x}}}\hat{v}(F_K^{-1}(\boldsymbol{x}))|^2 \le c\,\|\boldsymbol{T}_K\|^{2m} \sum_{|\alpha|=m} |\partial^{\alpha}_{\boldsymbol{x}} v(\boldsymbol{x})|^2,$$

and so

$$\begin{aligned} |\hat{v}|^2_{m,\hat{K}} &\le c \sum_{|\alpha|=m} \int_K \left[\partial^{\alpha}_{\boldsymbol{x}} v(\boldsymbol{x})\right]^2 \|\boldsymbol{T}_K\|^{2m} (\det \boldsymbol{T}_K)^{-1} dx \\ &= c\,\|\boldsymbol{T}_K\|^{2m} (\det \boldsymbol{T}_K)^{-1}\, |v|^2_{m,K}, \end{aligned}$$

from which the inequality (10.3.5) follows. □

We now combine the preceding theorems to obtain an estimate for the interpolation error in the semi-norm $|v - \Pi_K v|_{m,K}$.

Theorem 10.3.5 *Let k and m be nonnegative integers with $k > 0$, $k+1 \ge m$, and $\mathbb{P}_k(\hat{K}) \subset \hat{X}$. Let Π_K be the operators defined in* (10.3.2)*. Then there is a constant c depending only on $\hat{K}$ and $\hat{\Pi}$ such that*

$$|v - \Pi_K v|_{m,K} \le c\,\frac{h_K^{k+1}}{\rho_K^m} |v|_{k+1,K} \quad \forall\, v \in H^{k+1}(K). \tag{10.3.7}$$

Proof. From Theorem 10.3.1 we have $\hat{v} - \hat{\Pi}\hat{v} = (v - \Pi_K v) \circ F_K$. Consequently, using (10.3.6) we obtain

$$|v - \Pi_K v|_{m,K} \le c\,\|\boldsymbol{T}_K^{-1}\|^m |\det \boldsymbol{T}_K|^{1/2} |\hat{v} - \hat{\Pi}\hat{v}|_{m,\hat{K}}.$$

Using the estimate (10.3.3), we have

$$|v - \Pi_K v|_{m,K} \le c\,\|\boldsymbol{T}_K^{-1}\|^m |\det \boldsymbol{T}_K|^{1/2} |\hat{v}|_{k+1,\hat{K}}. \tag{10.3.8}$$

The inequality (10.3.5) with $m = k+1$ is

$$|\hat{v}|_{k+1,\hat{K}} \le c\,\|\boldsymbol{T}_K\|^{k+1} |\det \boldsymbol{T}_K|^{-1/2} |v|_{k+1,K}.$$

So from (10.3.8), we obtain

$$|v - \Pi_K v|_{m,K} \le c\,\|\boldsymbol{T}_K^{-1}\|^m \|\boldsymbol{T}_K\|^{k+1} |v|_{k+1,K}.$$

The estimate (10.3.7) now follows from an application of Lemma 10.2.2. □

The error estimate (10.3.7) is proved through the use of the reference element $\hat{K}$. The proof method can be termed the *reference element technique*. We notice that in the proof we only use the polynomial invariance

property (10.3.4) of the finite element interpolation on the reference element, and we do not need to use a corresponding polynomial invariance property on the real finite element. This feature is important when we analyze finite element spaces which are not based on affine-equivalent elements. For example, suppose the domain is partitioned into quadrilateral elements $\{K \mid K \in \mathcal{T}_h\}$. Then a reference element can be taken to be the unit square $\hat{K} = [0,1]^2$. For each element K, the mapping function F_K is bilinear, and it maps each vertex of the reference element $\hat{K}$ to a corresponding vertex of K. The first degree finite element space for approximating $V = H^1(\Omega)$ is

$$V_h = \{v_h \in C(\overline{\Omega}) \mid v_h \circ F_K \in \mathbb{Q}_{1,1}(\hat{K}),\ K \in \mathcal{T}_h\},$$

where

$$\mathbb{Q}_{1,1}(\hat{K}) = \{\hat{v} \mid \hat{v}(\hat{\boldsymbol{x}}) = a + b\,\hat{x}_1 + c\,\hat{x}_2 + d\,\hat{x}_1\hat{x}_2,\ a,b,c,d \in \mathbb{R}\}$$

is the space of bilinear functions. We see that for $v_h \in V_h$, on each element K, $v_h|_K$ is not necessarily a polynomial (as a function of the variable $\boldsymbol{x}$), but rather the image of a polynomial on the reference element. Obviously, we do not have the polynomial invariance property for the interpolation operator Π_K, nevertheless (10.3.4) is still valid. For such a finite element space, the proof of Theorem 10.3.5 still goes through.

The error bound in (10.3.7) depends on two parameters h_K and ρ_K. It will be convenient to use the parameter h_K only in an interpolation error bound. For this purpose we introduce the notion of a *regular* family of finite elements. For a triangulation $\mathcal{T}_h$, we denote

$$h = \max_{K \in \mathcal{T}_h} h_K, \tag{10.3.9}$$

often called the *mesh parameter*. The quantity h is a measure of how refined the mesh is. The smaller h is, the finer the mesh.

Definition 10.3.6 *A family $\{\mathcal{T}_h\}_h$ of finite element partitions is said to be* regular *if*
(a) *there exists a constant σ such that $h_K/\rho_K \le \sigma$ for all elements $K \in \mathcal{T}_h$ and for any h;*
(b) *the mesh parameter h approaches zero.*

A necessary and sufficient condition for the fulfillment of the condition (a) in Definition 10.3.6 is that the minimal angles of all the elements are bounded below away from 0; a proof of this result is left as an exercise (see Exercise 10.3.3).

In the case of a regular family of affine finite elements, we can deduce the following error estimate from Theorem 10.3.5.

Corollary 10.3.7 *We keep the assumptions stated in Theorem 10.3.5. Furthermore, assume $\{\mathcal{T}_h\}_h$ is a regular family of finite elements. Then there is a constant c such that for any $\mathcal{T}_h$ in the family,*

$$\|v-\Pi_K v\|_{m,K} \le c\, h_K^{k+1-m}|v|_{k+1,K} \quad \forall\, v \in H^{k+1}(K),\ \forall\, K \in \mathcal{T}_h. \tag{10.3.10}$$

Example 10.3.8 Let K be a triangle in a regular family of affine finite elements. We take the three vertices of K to be the nodal points. The local function space X_K is $\mathbb{P}_1(K)$. Assume $v \in H^2(K)$. Applying the estimate (10.3.10) with $k = 1$, we have

$$\|v - \Pi_K v\|_{m,K} \le c\, h_K^{2-m}|v|_{2,K} \quad \forall\, v \in H^2(K). \tag{10.3.11}$$

This estimate holds for $m = 0, 1$. □

10.3.4 Global interpolation error estimates

We now estimate the finite element interpolation error of a continuous function over the entire domain Ω. For a function $v \in C(\overline{\Omega})$, we construct its global interpolant $\Pi_h v$ in the finite element space X_h by the formula

$$\Pi_h v|_K = \Pi_K v \quad \forall\, K \in \mathcal{T}_h.$$

Let $\{\boldsymbol{x}_i\}_{i=1}^{N_h} \subset \overline{\Omega}$ be the set of the nodes collected from the nodes of all the elements $K \in \mathcal{T}_h$. We have the representation formula

$$\Pi_h v = \sum_{i=1}^{N_h} v(\boldsymbol{x}_i)\phi_i \tag{10.3.12}$$

for the global finite element interpolant. Here ϕ_i, $i = 1, \ldots, N_h$, are the global basis functions that span X_h. The basis function ϕ_i is associated with the node $\boldsymbol{x}_i$, i.e., ϕ_i is a piecewise polynomial of degree less than or equal to k, and $\phi_i(\boldsymbol{x}_j) = \delta_{ij}$. If the node $\boldsymbol{x}_i$ is a vertex $\boldsymbol{x}_l^K$ of the element K, then $\phi_i|_K = \phi_l^K$. If $\boldsymbol{x}_i$ is not a node of K, then $\phi_i|_K = 0$. Thus the functions ϕ_i are constructed from local basis functions ϕ_i^K.

In the context of the finite element approximation of a linear second-order elliptic boundary value problem, there holds the Céa's inequality (cf. (9.1.11))

$$\|u - u_h\|_{1,\Omega} \le c \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega}.$$

Then

$$\|u - u_h\|_{1,\Omega} \le c\, \|u - \Pi_h u\|_{1,\Omega},$$

and we need to find an estimate of the interpolation error $\|u - \Pi_h u\|_{1,\Omega}$.

Theorem 10.3.9 *Assume that all the conditions of Corollary 10.3.7 hold. Then there exists a constant c independent of h such that*

$$\|v - \Pi_h v\|_{m,\Omega} \le c\, h^{k+1-m}|v|_{k+1,\Omega} \quad \forall\, v \in H^{k+1}(\Omega),\ m = 0, 1. \tag{10.3.13}$$

Proof. Since the finite element interpolant $\Pi_h u$ is defined piecewisely by $\Pi_h u|_K = \Pi_K u$, we can apply Corollary 10.3.7 with $m = 0$ and 1 to find

$$\begin{aligned}\|u - \Pi_h u\|_{m,\Omega}^2 &= \sum_{K\in\mathcal{T}_h} \|u - \Pi_K u\|_{m,K}^2 \\ &\le \sum_{K\in\mathcal{T}_h} c\, h_K^{2(k+1-m)} |u|_{k+1,K}^2 \\ &\le c\, h^{2(k+1-m)} |u|_{k+1,\Omega}^2.\end{aligned}$$

Taking the square root of the above relation, we obtain the error estimates (10.3.13). □

We make a remark on finite element interpolation of possibly discontinuous functions. The finite element interpolation error analysis discussed above assumes the function to be interpolated is continuous, so that it is meaningful to talk about its finite element interpolation (10.3.12). In the case of a general Sobolev function v, not necessarily continuous, we can define $\Pi_h v$ by local L^2 projections in such a way that the interpolation error estimates stated in Theorem 10.3.9 are still valid. For detail, see [54]. We will use the same symbol $\Pi_h v$ to denote the "regular" finite element interpolant (10.3.12) when v is continuous, and in case v is discontinuous, $\Pi_h v$ is defined through local L^2 projections. In either case, we have the error estimates (10.3.13).

Exercise 10.3.1 Let $\{\mathcal{T}_h\}_h$ be a regular family of finite element partitions of a polygon $\overline{\Omega}$ into triangles. For each $\mathcal{T}_h$, let

$$Q_h = \left\{ v_h \in L^2(\Omega) \mid v_h|_K \in \mathbb{R}\ \forall\, K \in \mathcal{T}_h \right\},$$

and define a piecewise averaging operator $P_h : L^2(\Omega) \to Q_h$ by

$$(P_h v)|_K = \int_K v\, dx \Big/ \mathrm{meas}(K) \quad \forall\, K \in \mathcal{T}_h.$$

Use the reference element technique to prove the following error bound:

$$\|v - P_h v\|_{L^2(\Omega)} \le c\, h\, |v|_{H^1(\Omega)} \quad \forall\, v \in H^1(\Omega).$$

Moreover, show the following:

$$\left| \int_\Omega u\,(v - P_h v)\, dx \right| \le c\, h^2 |u|_{H^1(\Omega)} |v|_{H^1(\Omega)} \quad \forall\, u, v \in H^1(\Omega).$$

Exercise 10.3.2 Let $\{\mathcal{T}_h\}_h$ be a regular family of partitions of the domain $\overline{\Omega}$, and let any element K be obtained from the reference element $\hat{K}$ through the affine mapping (10.2.25). Let $1 \le p < \infty$. Show that there exists a constant c independent of K and h such that

$$\int_{\partial K} |v|^p ds \le c \left(h_K^{-1} \|v\|_{0,p,K}^p + h_K^{p-1} |v|_{1,p,K}^p \right) \quad \forall\, v \in W^{1,p}(K).$$

In particular, with $p = 2$,

$$\int_{\partial K} |v|^2 ds \le c \left(h_K^{-1} \|v\|_{0,K}^2 + h_K |v|_{1,K}^2 \right) \quad \forall\, v \in H^1(K).$$

Exercise 10.3.3 Show that a necessary and sufficient condition for requirement (a) in Definition 10.3.6 is that the minimal angles of all the elements are bounded below from 0.

Exercise 10.3.4 A family of triangulations $\{\mathcal{T}_h\}$ is said to be *quasiuniform* if the family is regular, and there is a constant $c_0 > 0$ such that

$$\min_{K \in \mathcal{T}_h} h_K / \max_{K \in \mathcal{T}_h} h_K \ge c_0 \quad \forall\, \mathcal{T}_h.$$

(Hence, all the elements in $\mathcal{T}_h$ are of comparable size.)

Suppose $\{\mathcal{T}_h\}$ is a family of quasiuniform triangulations of the domain $\Omega \subset \mathbb{R}^2$, and $\{X_h\}$ is a corresponding family of affine-equivalent finite elements. Denote $N_h = \dim X_h$, and denote the nodes (of the basis functions) by $\boldsymbol{x}_i$, $1 \le i \le N_h$. Show that there are constants $c_1, c_2 > 0$ independent of h such that

$$c_1 \|v\|_{L^2(\Omega)}^2 \le h^2 \sum_{i=1}^{N_h} |v(\boldsymbol{x}_i)|^2 \le c_2 \|v\|_{L^2(\Omega)}^2 \quad \forall\, v \in X_h.$$

Exercise 10.3.5 The Bramble-Hilbert Lemma ([40]) states: Let $\Omega_0 \subset \mathbb{R}^d$ be a Lipschitz domain, and let $\ell : H^{k+1}(\Omega_0) \to \mathbb{R}$ be bounded and satisfy

$$|\ell(v_1 + v_2)| \le |\ell(v_1)| + |\ell(v_2)| \quad \forall\, v_1, v_2 \in H^{k+1}(\Omega_0).$$

Suppose $\ell(v) = 0\ \forall\, v \in \mathbb{P}_k(\Omega_0)$. Then there exists a constant c, depending on Ω_0, such that

$$|\ell(v)| \le c\, |v|_{k+1,\Omega_0} \quad \forall\, v \in H^{k+1}(\Omega_0).$$

(a) Prove the Bramble-Hilbert Lemma.
(b) Deduce Theorem 10.3.3 from the Bramble-Hilbert Lemma.

Exercise 10.3.6 Prove the following extension of the Bramble-Hilbert Lemma: Let $\Omega_0 \subset \mathbb{R}^d$ be a Lipschitz domain, and let $a(\cdot,\cdot) : H^{k+1}(\Omega_0) \times H^{l+1}(\Omega_0) \to \mathbb{R}$ be a bounded functional such that

$$|a(u_1 + u_2, v)| \le |a(u_1, v)| + |a(u_2, v)| \quad \forall\, u_1, u_2 \in H^{k+1}(\Omega_0),\ v \in H^{l+1}(\Omega_0),$$
$$|a(u, v_1 + v_2)| \le |a(u, v_1)| + |a(u, v_2)| \quad \forall\, u \in H^{k+1}(\Omega_0),\ v_1, v_2 \in H^{l+1}(\Omega_0).$$

Assume

$$a(u, v) = 0 \quad \forall\, u \in H^{k+1}(\Omega_0),\ v \in \mathbb{P}_l(\Omega_0),$$
$$a(u, v) = 0 \quad \forall\, u \in \mathbb{P}_k(\Omega_0),\ v \in H^{l+1}(\Omega_0).$$

Then there exists a constant c, depending on Ω_0, such that

$$|a(u, v)| \le c\, |u|_{k+1,\Omega_0} |v|_{l+1,\Omega_0} \quad \forall\, u \in H^{k+1}(\Omega_0),\ v \in H^{l+1}(\Omega_0).$$

Exercise 10.3.7 In this exercise, we employ the technique of the reference element to estimate the condition number of the stiffness matrix. The boundary value problem considered is a symmetric elliptic second-order problem

$$u \in V, \quad a(u,v) = \ell(v) \quad \forall\, v \in V,$$

where $V \subset H^1(\Omega)$ is a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is bilinear, symmetric, continuous and V-elliptic, $\ell \in V'$. Let $V_h \subset V$ be an affine family of finite elements of piecewiese polynomials of degree less than or equal to r. The finite element solution $u_h \in V_h$ is defined by

$$u_h \in V_h, \quad a(u_h, v_h) = \ell(v_h) \quad \forall\, v_h \in V_h.$$

Let $\{\phi_i\}$ be a basis of the space V_h. If we express the finite element solution in terms of the basis, $u_h = \sum_i \xi_i \phi_i$, then the unknown coefficients are determined from a linear system

$$A\boldsymbol{\xi} = \boldsymbol{b},$$

where the stiffness matrix A has the entries $a(\phi_j, \phi_i)$. Then A is a symmetric positive definite matrix. Let us find an upper bound for the spectral condition number

$$\mathrm{Cond}_2(A) = \|A\|_2 \|A^{-1}\|_2.$$

We assume the finite element spaces $\{V_h\}$ are constructed based on a family of quasiuniform triangulations $\{\mathcal{T}_h\}$.
(1) Show that there exist constants $c_1, c_2 > 0$, such that

$$c_1 h^2 |\boldsymbol{\eta}|^2 \le \|v_h\|_0^2 \le c_2 h^2 |\boldsymbol{\eta}|^2 \quad \forall\, v_h \in V_h,\ v_h = \sum_i \eta_i \phi_i.$$

(2) Show that there exists a constant $c_3 > 0$ such that

$$\|\nabla v_h\|_0^2 \le c_3 h^{-2} \|v_h\|_0^2 \quad \forall\, v_h \in V_h.$$

This result is an example of an inverse inequality for finite element functions.
(3) Show that $\mathrm{Cond}_2(A) = \mathcal{O}(h^{-2})$.
Hint: Since A is symmetric, positive definite, $\|A\|_2 = \sup\{(A\boldsymbol{\eta}, \boldsymbol{\eta})/|\boldsymbol{\eta}|^2\}$.

In the general d-dimensional case, it can be shown that these results are valid with the h^2 terms in (1) being replaced by h^d; in particular, we notice that the result (3) does not depend on the dimension of the domain Ω.

10.4 Convergence and error estimates

As an example, we consider the convergence and error estimates for finite element approximations of a linear second-order elliptic problem over a polygonal domain. The function space V is a subspace of $H^1(\Omega)$; e.g., $V = H_0^1(\Omega)$ if the homogeneous Dirichlet condition is specified over the whole boundary, whereas $V = H^1(\Omega)$ if a Neumann condition is specified over the boundary. Let the weak formulation of the problem be

$$u \in V, \quad a(u,v) = \ell(v) \quad \forall\, v \in V. \tag{10.4.1}$$

We assume all the assumptions required by the Lax-Milgram Lemma; then the problem (10.3.13) has a unique solution u. Let $V_h \subset V$ be a finite element space. Then the discrete problem

$$u_h \in V_h, \quad a(u_h, v_h) = \ell(v_h) \quad \forall\, v_h \in V_h \tag{10.4.2}$$

also has a unique solution $u_h \in V_h$ and there holds the Céa's inequality

$$\|u - u_h\|_V \le c \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{10.4.3}$$

This inequality is a basis for convergence and error analysis.

Theorem 10.4.1 *We keep the assumptions mentioned above. Let $k > 0$ be an integer, and let $\{V_h\} \subset V$ be affine-equivalent finite element spaces of piecewise polynomials of degree less than or equal to k, corresponding to a regular family of triangulations of $\overline{\Omega}$. Then the finite element method converges:*

$$\|u - u_h\|_V \to 0 \quad \text{as } h \to 0.$$

Assume $u \in H^{k+1}(\Omega)$. Then there exists a constant c such that the following error estimate holds:

$$\|u - u_h\|_{1,\Omega} \le c\, h^k |u|_{k+1,\Omega}. \tag{10.4.4}$$

Proof. We take $v_h = \Pi_h u$ in Céa's inequality (10.4.3),

$$\|u - u_h\|_{1,\Omega} \le c\, \|u - \Pi_h u\|_{1,\Omega}.$$

Using the estimate (10.3.13) with $m = 1$, we obtain the error estimate (10.4.4).

The convergence of the finite element solution under the basic solution regularity $u \in V$ follows from the facts that smooth functions are dense in the space V and for a smooth function, its finite element interpolants converge (with a convergence order k). □

Example 10.4.2 Consider the problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma. \end{aligned}$$

The corresponding variational formulation is: Find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall\, v \in H_0^1(\Omega).$$

This problem has a unique solution. Similarly, the discrete problem of finding $u_h \in V_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h dx = \int_\Omega f v_h dx \quad \forall\, v_h \in V_h$$

has a unique solution. Here V_h is a finite element subspace of $H_0^1(\Omega)$, consisting of piecewise polynomials of degree less than or equal to k, corresponding to a regular triangulation of $\overline{\Omega}$. If $u \in H^{k+1}(\Omega)$, then the error is estimated by

$$\|u - u_h\|_{1,\Omega} \leq c\, h^k \|u\|_{k+1,\Omega}.$$

□

It is possible to derive error estimates for the finite element solutions in other norms, such as the $L^2(\Omega)$ norm or $L^\infty(\Omega)$ norm. In the following, we show how to derive estimates for the $L^2(\Omega)$ norm error $\|u - u_h\|_{0,\Omega}$.

Theorem 10.4.3 (Aubin-Nitsche Lemma) *In the context of Theorem 10.4.1, we have*

$$\|u - u_h\|_{0,\Omega} \leq M \|u - u_h\|_{1,\Omega} \sup_{g \in L^2(\Omega)} \left(\frac{1}{\|g\|_{0,\Omega}} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_{1,\Omega} \right), \tag{10.4.5}$$

where for each $g \in L^2(\Omega)$, $\varphi_g \in V$ is the solution of the problem

$$a(v, \varphi_g) = (g, v)_{0,\Omega} \quad \forall\, v \in V. \tag{10.4.6}$$

Proof. First note that under the assumptions, (10.4.6) has a unique solution.

Next, denote $e = u - u_h$ and recall the error relation

$$a(e, v_h) = 0 \quad \forall\, v_h \in V_h. \tag{10.4.7}$$

Let $\varphi_e \in V$ be the solution of the problem (10.4.6) with $g = e$:

$$a(v, \varphi_e) = (e, v)_{0,\Omega} \quad \forall\, v \in V. \tag{10.4.8}$$

Take $v = e$ in (10.4.8) and use (10.4.7) to obtain

$$\|e\|_{0,\Omega}^2 = a(e, \varphi_e) = a(e, \varphi_e - v_h) \quad \forall\, v_h \in V_h.$$

Then,

$$\|e\|_{0,\Omega}^2 \leq M \|e\|_{1,\Omega} \inf_{v_h \in V_h} \|\varphi_e - v_h\|_{1,\Omega}.$$

If $\|e\|_{0,\Omega} = 0$, then (10.4.5) is obvious. Otherwise, we rewrite the above inequality as

$$\|e\|_{0,\Omega} \leq M \|e\|_{1,\Omega} \frac{1}{\|e\|_{0,\Omega}} \inf_{v_h \in V_h} \|\varphi_e - v_h\|_{1,\Omega}.$$

Therefore, the relations (10.4.5) holds. □

The term

$$\sup_{g \in L^2(\Omega)} \left(\frac{1}{\|g\|_{0,\Omega}} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_{1,\Omega} \right)$$

in (10.4.5) is usually small.

Corollary 10.4.4 *Suppose the solution $\varphi_g \in V$ of the problem* (10.4.6) *has the regularity $\varphi_g \in H^2(\Omega)$ and the following regularity bound holds:*

$$\|\varphi_g\|_{2,\Omega} \le c\,\|g\|_{0,\Omega} \quad \forall\, g \in L^2(\Omega). \tag{10.4.9}$$

Then we have the following inequality from (10.4.5)*:*

$$\|u - u_h\|_{0,\Omega} \le c\,h\,\|u - u_h\|_{1,\Omega}. \tag{10.4.10}$$

Proof. Since $\varphi_g \in H^2(\Omega)$, we have

$$\inf_{v_h \in V_h} \|\varphi_g - v_h\|_{1,\Omega} \le c\,h\,|\varphi_g|_{2,\Omega}.$$

The term $|\varphi_g|_{2,\Omega}$ is bounded by $c\,\|g\|_{0,\Omega}$, by the regularity bound (10.4.9). Hence,

$$\sup_{g \in L^2(\Omega)} \left(\frac{1}{\|g\|_{0,\Omega}} \inf_{v_h \in V_h} \|\varphi_g - v_h\|_{1,\Omega} \right) \le c\,h$$

and the inequality (10.4.10) holds. □

Combining (10.4.10) and (10.4.4), we conclude that under the assumptions stated in Theorem 10.4.1 and Corollary 10.4.4, we have the following optimal order error estimate in $L^2(\Omega)$ norm:

$$\|u - u_h\|_{0,\Omega} \le c\,h^{k+1}|u|_{k+1,\Omega}. \tag{10.4.11}$$

A crucial condition for the relation (10.4.10) is the solution regularity bound (10.4.9), its validity depending not only on the smoothness of the coefficients, the right hand side, and the boundary condition functions of the boundary value problem, but also on the smoothness of the domain, as well as whether the type of the boundary condition changes. For example, consider the boundary value problem of the Poisson equation in a domain $\Omega \subset \mathbb{R}^2$:

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

If Ω is smooth, say $\partial\Omega \in C^{1,1}$, or Ω is convex, then (10.4.9) holds. However, if Ω is a non-convex polygon or if the boundary condition type changes at a boundary point (e.g., from Dirichlet to Neumann), then the solution has singularities and the solution regularity bound (10.4.9) does not hold. Detailed discussion of this topic can be found in [98]. Exercise 10.4.6 provides several examples of singularities caused by the boundary corners or change of boundary condition type.

We emphasize that in error estimation for finite element solutions, we need to assume certain degree of solution regularity. When the solution exhibits singularities, there are two popular approaches to recover the optimal

convergence order corresponding to a smooth solution. One approach is by means of *singular elements*, i.e., some singular functions are included in the finite element space. The advantage of this approach is its efficiency, while the weakness is that the form of the singular functions must be known *a priori*. The second approach is by using *mesh refinement* around the singularities. This approach does not need the knowledge on the forms of the singular functions, and is more popular in practical use.

Exercise 10.4.1 Let us use linear elements to solve the boundary value problem

$$\begin{cases} -u'' = f \quad \text{in } (0,1), \\ u(0) = u(1) = 0, \end{cases}$$

where $f \in L^2(0,1)$. Divide the domain $\overline{\Omega} = [0,1]$ with the nodes $0 = x_0 < x_1 < \cdots < x_N = 1$, and denote the elements $K_i = [x_{i-1}, x_i]$, $1 \le i \le N$. Then the finite element space is

$$V_h = \{v_h \in H_0^1(0,1) \mid v_h|_{I_i} \in \mathbb{P}_1(K_i),\ 1 \le i \le N\}.$$

Let $u_h \in V_h$ denote the corresponding finite element solution of the boundary value problem. Prove that $u_h(x_i) = u(x_i)$, $0 \le i \le N$; in other words, the linear finite element solution for the boundary value problem is infinitely accurate at the nodes.
Hint: Show that the finite element interpolant of u is the finite element solution.

Exercise 10.4.2 Let $\Omega = (0,1)^2$ and $f \in L^2(\Omega)$. For the following BVP

$$-\left(\frac{\partial^2 u}{\partial x_1^2} + 2\,\frac{\partial^2 u}{\partial x_2^2}\right) + \sin(x_1 x_2)\, u = f \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } \partial\Omega,$$

develop a finite element method, and provide error bounds in $H^1(\Omega)$-norm and $L^2(\Omega)$-norm.

Exercise 10.4.3 In Exercise 8.5.2, we studied the weak formulation of an elasticity problem for an isotropic, homogeneous linearly elastic material. Let $d = 2$ and assume Ω is a polygonal domain. Introduce a regular family of finite element partitions $\mathcal{T}_h = \{K\}$ in such a way that each K is a triangle and if $K \cap \Gamma \neq \emptyset$, then either $K \cap \Gamma \subset \Gamma_D$ or $K \cap \Gamma \subset \overline{\Gamma_N}$. Let $V_h \subset V$ be the corresponding finite element subspace of continuous piecewise linear functions. Give the formulation of the finite element method and show that there is a unique finite element solution $\boldsymbol{u}_h \in V_h$.

Assume $\boldsymbol{u} \in (H^2(\Omega))^2$. Derive error estimates for $\boldsymbol{u} - \boldsymbol{u}_h$ and $\boldsymbol{\sigma} - \boldsymbol{\sigma}_h$, where

$$\boldsymbol{\sigma}_h = \lambda \operatorname{tr} \boldsymbol{\varepsilon}(\boldsymbol{u}_h)\, \boldsymbol{I} + 2\,\mu\, \boldsymbol{\varepsilon}(\boldsymbol{u}_h)$$

is a discrete stress field.

Exercise 10.4.4 Consider a finite element approximation of the nonlinear elliptic boundary value problem studied in Section 8.8. Let us use all the notation introduced there. Let V_h be a finite element space consisting of continuous, piecewise polynomials of certain degree such that the functions vanish on the boundary. Then from Example 7.2.7, $V_h \subset V$. Show that the finite element method

$$u_h \in V_h, \quad a(u_h; u_h, v_h) = \ell(v_h) \quad \forall\, v_h \in V_h$$

has a unique solution. Also show that u_h is the unique minimizer of the energy functional $E(\cdot)$ over the finite element space V_h.

Error estimate for the finite element solution defined above can be derived following that in [52, Section 5.3], where finite element approximation of the homogeneous Dirichlet problem for the nonlinear differential equation

$$-\text{div}\,(|\nabla u|^{p-2}\nabla u) = f$$

is considered. However, the error estimate is not of optimal order. The optimal order error estimate for the linear element solution is derived in [32].

Exercise 10.4.5 Show that in $\mathbb{R}^2$, in terms of the polar coordinates

$$x_1 = r\,\cos\theta, \quad x_2 = r\,\sin\theta,$$

the Laplacian operator takes the form

$$\begin{aligned}\Delta &= \frac{\partial^2}{\partial r^2} + \frac{1}{r}\,\frac{\partial}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2}{\partial \theta^2} \\ &= \frac{1}{r}\,\frac{\partial}{\partial r}\left(r\,\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\,\frac{\partial^2}{\partial \theta^2};\end{aligned}$$

and in $\mathbb{R}^3$, in terms of the spherical coordinates

$$x_1 = r\,\cos\theta\,\sin\phi, \quad x_2 = r\,\sin\theta\,\sin\phi, \quad x_3 = r\,\cos\phi,$$

the Laplacian operator takes the form

$$\begin{aligned}\Delta &= \frac{\partial^2}{\partial r^2} + \frac{2}{r}\,\frac{\partial}{\partial r} + \frac{1}{r^2}\left(\frac{1}{\sin^2\phi}\,\frac{\partial^2}{\partial \theta^2} + \cot\phi\frac{\partial}{\partial \phi} + \frac{\partial^2}{\partial \phi^2}\right) \\ &= \frac{1}{r^2}\,\frac{\partial}{\partial r}\left(r^2\,\frac{\partial}{\partial r}\right) + \frac{1}{r^2}\left(\frac{1}{\sin^2\phi}\,\frac{\partial^2}{\partial \theta^2} + \cot\phi\frac{\partial}{\partial \phi} + \frac{\partial^2}{\partial \phi^2}\right).\end{aligned}$$

Exercise 10.4.6 Consider the domain $\Omega = \{(r,\theta) \mid 0 < r < 1,\, 0 < \theta < \omega\}$, $\omega \le 2\pi$. Its boundary consists of two straight sides

$$\begin{aligned}\Gamma_1 &= \{(r,\theta) \mid 0 \le r \le 1,\, \theta = 0\}, \\ \Gamma_2 &= \{(r,\theta) \mid 0 \le r \le 1,\, \theta = \omega\}\end{aligned}$$

and one circular curve

$$\Gamma_3 = \{(r,\theta) \mid r = 1,\, 0 \le \theta \le \omega\}.$$

Denote $\alpha = \pi/\omega$.
(a) Show that $u = r^\alpha \sin(\alpha\theta)$ is the solution of the boundary value problem

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_1 \cup \Gamma_2, \\ u &= \sin(\alpha\theta) \quad \text{on } \Gamma_3. \end{aligned}$$

Verify that $u \in H^1(\Omega)$, and when $\omega > \pi$, $u \notin H^2(\Omega)$.
(b) Show that $u = r^\alpha \cos(\alpha\theta)$ is the solution of the boundary value problem

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega, \\ \partial u/\partial \nu &= 0 \quad \text{on } \Gamma_1 \cup \Gamma_2, \\ u &= \cos(\alpha\theta) \quad \text{on } \Gamma_3, \end{aligned}$$

where $\partial/\partial\nu$ is the outward normal derivative. Verify that $u \in H^1(\Omega)$, and when $\omega > \pi$, $u \notin H^2(\Omega)$.
(c) Show that $u = r^{\alpha/2} \sin(\alpha\theta/2)$ is the solution of the boundary value problem

$$\begin{aligned} -\Delta u &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_1, \\ \partial u/\partial \nu &= 0 \quad \text{on } \Gamma_2, \\ u &= \sin(\alpha\theta/2) \quad \text{on } \Gamma_3. \end{aligned}$$

Verify that $u \in H^1(\Omega)$, and when $\omega > \pi/2$, $u \notin H^2(\Omega)$. Observe that due to the boundary condition type change at the boundary point $(0, 0)$, the strength of the solution singularity doubles.

Suggestion for Further Reading.

Some standard references on mathematical analysis of the finite element method include Babuška and Aziz [29], Braess [39], Brenner and Scott [42], Ciarlet [52, 53], Johnson [131], Oden and Reddy [183], Strang and Fix [216].

Convergence of the finite element method may be achieved by progressively refining the mesh, or by increasing the polynomial degree, or by doing both simultaneously. Then we get the h-version, p-version, or h-p-version of the finite element method. It is customary to use h as the parameter for the meshsize and p as the parameter for the polynomial degree. Efficient selection among the three versions of the method depends on the *a priori* knowledge on the regularity of the exact solution of the problem. Roughly speaking, over a region where the solution is smooth, high degree polynomials with large size elements are more efficient, and in a region where the solution has singularities, low order elements together with a locally refined mesh should be used. Detailed discussion of the p-version finite element method can be found in Szabó and Babuška [219]. Mathematical theory of the p-version and h-p-version finite element methods with applications in solid and fluid mechanics can be found in Schwab [204]. The

recent monograph Šolín, Segeth and Doležel [209] provides detailed description of practical implementation of the h-p-version finite element methods.

For the theory of mixed and hybrid finite element methods, see Brezzi and Fortin [43], Roberts and Thomas [196].

For the numerical solution of Navier-Stokes equations by the finite element method, see Girault and Raviart [90].

Theory of the finite element method for solving parabolic problems can be found in Thomée [222].

Singularities of solutions to boundary value problems on non-smooth domains are analyzed in detail in Grisvard [98]. See also Kozlov, Maz'ya and Rossmann [146]. To improve the convergence rate of the finite element method when the solution exhibits singularities, one can employ the so-called singular element method where the finite element space contains the singular functions, or the mesh refinement method where the mesh is locally refined around the singular region of the solution. One can find a discussion of the singular element method in Strang and Fix [216], and the mesh refinement method in Szabó and Babuška [219].

Adaptive finite element methods based on a posteriori error estimates have attracted much attention for nearly 30 years. Two comprehensive references on this subject are Ainsworth and Oden [2] and Babuška and Strouboulis [30].

Discontinuous Galerkin methods have been a very active research topic in the past two decades. The methods use discontinuous approximations, and thus have various advantages over the standard finite element method, such as handling easily complicated geometry and different polynomial degrees in different elements. A general reference on this topic is a volume edited by Cockburn, Karniadakis, and Shu [55]. A unified error analysis of the methods for second-order elliptic problems is given in a paper by Arnold, Brezzi, Cockburn, and Marini [8].

General discussions on solving linear systems can be found in Atkinson [15, Chap. 8], Golub and Van Loan [95], Stewart [214]. A very efficient way of solving finite element systems is through multi-grid methods, see e.g., Hackbusch [104], Xu [235].

11

Elliptic Variational Inequalities and Their Numerical Approximations

Variational inequalities form an important family of nonlinear problems. Some of the more complex physical processes are described by variational inequalities. We study some elliptic variational inequalities (EVIs) in this chapter, presenting results on existence, uniqueness and stability of solutions to the EVIs, and discussing their numerical approximations.

We start with an introduction of two sample elliptic variational inequalities (EVIs) in Section 11.1. Many elliptic variational inequalities arising in mechanics can be equivalently expressed as convex minimization problems; a study of such variational inequalities is given in Section 11.2. Then in Section 11.3, we consider a general family of EVIs that do not necessarily relate to minimization problems. Numerical approximations of the EVIs are the topics of Section 11.4. In the last section, we consider some contact problems in elasticity that lead to EVIs.

11.1 From variational equations to variational inequalities

As in previous chapters, we assume $\Omega \subset \mathbb{R}^d$ is a Lipschitz domain so that the unit outward normal vector $\boldsymbol{\nu}$ exists a.e. on the boundary Γ.

Consider the model elliptic boundary value problem:

$$-\Delta u = f \quad \text{in } \Omega, \tag{11.1.1}$$

$$u = 0 \quad \text{on } \Gamma, \tag{11.1.2}$$

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_11,
© Springer Science + Business Media, LLC 2009

where $f \in L^2(\Omega)$ is given. Its weak formulation is

$$u \in H_0^1(\Omega) \quad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f \, v \, dx \quad \forall \, v \in H_0^1(\Omega). \tag{11.1.3}$$

By the Lax-Milgram Lemma, the problem (11.1.3) has a unique solution. Moreover, the problem (11.1.3) is equivalent to the minimization problem

$$u \in V, \quad E(u) = \inf_{v \in V} E(v), \tag{11.1.4}$$

where $V = H_0^1(\Omega)$ and

$$E(v) = \int_\Omega \left(\frac{1}{2} |\nabla v|^2 - f \, v \right) dx. \tag{11.1.5}$$

We observe that a minimization problem of the form (11.1.4) is a *linear* problem, owing to the properties that $E(\cdot)$ is a quadratic functional, and the set V over which the infimum is sought is a linear space. The problem (11.1.4) becomes nonlinear if the energy functional $E(v)$ is no longer quadratic (in particular, if $E(v)$ contains a non-differentiable term), or the energy functional is minimized over a general (convex) set instead of a linear space, or both.

Let us examine two concrete examples.

Example 11.1.1 (Obstacle problem) In an obstacle problem, we want to determine the equilibrium position of an elastic membrane which (1) passes through a closed curve Γ, the boundary of a planar domain Ω; (2) lies above an obstacle of height ψ; and (3) is subject to the action of a vertical force of density τf, here τ is the elastic tension of the membrane, and f is a given function.

Denote by u the vertical displacement component of the membrane. Since the membrane is fixed along the boundary Γ, we have the boundary condition $u = 0$ on Γ. To make the problem meaningful, we assume the obstacle function satisfies the condition $\psi \leq 0$ on Γ. In the following, we assume $\psi \in H^1(\Omega)$ and $f \in L^2(\Omega)$. Then the set of admissible displacements is

$$K = \{ v \in H_0^1(\Omega) \mid v \geq \psi \text{ a.e. in } \Omega \}.$$

The principle of minimal energy from mechanics asserts that the displacement u is a minimizer of the (scaled) total energy:

$$u \in K, \quad E(u) = \inf_{v \in K} E(v), \tag{11.1.6}$$

where the energy functional is defined in (11.1.5). It is easy to show that the solution is also characterized by the variational inequality (Exercise 11.1.1)

$$u \in K, \quad \int_\Omega \nabla u \cdot \nabla (v - u) \, dx \geq \int_\Omega f (v - u) \, dx \quad \forall \, v \in K. \tag{11.1.7}$$

It is possible to derive a classical form of pointwise relations for the variational inequality (11.1.7). For this purpose, we assume additionally that

$$f \in C(\Omega), \quad \psi \in C(\Omega), \quad u \in C^2(\Omega) \cap C(\overline{\Omega}). \tag{11.1.8}$$

We take $v = u + \phi$ in (11.1.7), with $\phi \in C_0^\infty(\Omega)$ and $\phi \geq 0$ in Ω, to obtain

$$\int_\Omega (\nabla u \cdot \nabla \phi - f\,\phi)\,dx \geq 0.$$

Perform an integration by parts,

$$\int_\Omega (-\Delta u - f)\,\phi\,dx \geq 0 \quad \forall\, \phi \in C_0^\infty(\Omega),\ \phi \geq 0 \text{ in } \Omega.$$

We see then that u must satisfy the differential inequality

$$-\Delta u - f \geq 0 \quad \text{in } \Omega.$$

Now suppose for some $\boldsymbol{x}_0 \in \Omega$, $u(\boldsymbol{x}_0) > \psi(\boldsymbol{x}_0)$. Then there exist a neighborhood $U(\boldsymbol{x}_0) \subset \Omega$ of $\boldsymbol{x}_0$ and a number $\delta > 0$ such that $u(\boldsymbol{x}) > \psi(\boldsymbol{x}) + \delta$ for $\boldsymbol{x} \in U(\boldsymbol{x}_0)$. We use the symbol $C_0^\infty(U(\boldsymbol{x}_0))$ to denote a subspace of $C_0^\infty(\Omega)$ in which each function has a support contained in $U(\boldsymbol{x}_0)$. In (11.1.7) we choose $v = u \pm \delta\,\phi$ with any $\phi \in C_0^\infty(U(\boldsymbol{x}_0))$ satisfying $\|\phi\|_\infty \leq 1$ and perform an integration by parts to obtain the relation

$$\pm \int_\Omega (-\Delta u - f)\,\phi\,dx \geq 0 \quad \forall\, \phi \in C_0^\infty(U(\boldsymbol{x}_0)),\ \|\phi\|_\infty \leq 1.$$

Therefore,

$$\int_\Omega (-\Delta u - f)\,\phi\,dx = 0 \quad \forall\, \phi \in C_0^\infty(U(\boldsymbol{x}_0)),\ \|\phi\|_\infty \leq 1$$

and then

$$\int_\Omega (-\Delta u - f)\,\phi\,dx = 0 \quad \forall\, \phi \in C_0^\infty(U(\boldsymbol{x}_0)).$$

Hence, if $u(\boldsymbol{x}_0) > \psi(\boldsymbol{x}_0)$ and $\boldsymbol{x}_0 \in \Omega$, then

$$-\Delta u - f = 0 \quad \text{in } U(\boldsymbol{x}_0)$$

and in particular,

$$(-\Delta u - f)(\boldsymbol{x}_0) = 0.$$

Summarizing, under the additional regularity assumption (11.1.8), the following relations hold in Ω:

$$u - \psi \geq 0, \quad -\Delta u - f \geq 0, \quad (u - \psi)(-\Delta u - f) = 0. \tag{11.1.9}$$

Consequently, the domain Ω is decomposed into two parts. On one part, denoted by Ω_1, we have

$$u > \psi \quad \text{and} \quad -\Delta u - f = 0 \quad \text{in } \Omega_1,$$

and the membrane has no contact with the obstacle. On the remaining part, denoted by Ω_2, we have

$$u = \psi \ \text{ and } \ -\Delta u - f \geq 0 \ \text{ in } \Omega_2,$$

and there is contact between the membrane and the obstacle. Notice that the region of contact, $\{\boldsymbol{x} \in \Omega \mid u(\boldsymbol{x}) = \psi(\boldsymbol{x})\}$, is an unknown a priori. Because of this, the obstacle problem is a free-boundary problem.

A slight modification of the above argument shows that under the additional assumptions

$$f \in L^2(\Omega), \quad \psi \in C(\Omega), \quad u \in H^2(\Omega) \cap C(\overline{\Omega}), \tag{11.1.10}$$

we have

$$u - \psi \geq 0, \quad -\Delta u - f \geq 0, \quad (u - \psi)(-\Delta u - f) = 0 \quad \text{a.e. in } \Omega. \tag{11.1.11}$$

Next, let us show that if $u \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfies the pointwise relations (11.1.9) a.e. in Ω, then u must be a solution of the variational inequality (11.1.7). First we have

$$\int_\Omega (-\Delta u - f)\,(v - \psi)\,dx \geq 0 \quad \forall\, v \in K.$$

Since

$$\int_\Omega (-\Delta u - f)\,(u - \psi)\,dx = 0,$$

we obtain

$$\int_\Omega (-\Delta u - f)\,(v - u)\,dx \geq 0 \quad \forall\, v \in K.$$

Integrating by parts, we get

$$\int_\Omega \nabla u \cdot \nabla (v - u)\,dx \geq \int_\Omega f\,(v - u)\,dx \quad \forall\, v \in K.$$

Thus, u is a solution of the variational inequality (11.1.7).

In this example, we obtain a variational inequality from minimizing a quadratic energy functional over a convex set. □

Example 11.1.2 In the second example, we consider a problem of the form (11.1.4) with $V = H^1(\Omega)$ and

$$E(v) = \int_\Omega \left[\frac{1}{2}\left(|\nabla v|^2 + v^2\right) - fv\right] dx + g \int_\Gamma |v|\,ds,$$

where $g > 0$ and $f \in L^2(\Omega)$ are given. It is left as an exercise (Exercise 11.1.2) to show that the minimization problem

$$u \in V, \quad E(u) = \inf_{v \in V} E(v) \tag{11.1.12}$$

is equivalent to the variational inequality

$$u \in V, \quad \int_\Omega \left[\nabla u \cdot \nabla(v-u) + u\,(v-u)\right] dx + g \int_\Gamma \left(|v| - |u|\right) ds \\ \geq \int_\Omega f\,(v-u)\, ds \quad \forall\, v \in V. \tag{11.1.13}$$

A derivation similar to the one used in Example 11.1.1 shows that the corresponding pointwise formulation is

$$-\Delta u + u = f \quad \text{in } \Omega, \tag{11.1.14}$$

$$\left|\frac{\partial u}{\partial \nu}\right| \leq g, \ \frac{\partial u}{\partial \nu}\, u + g\,|u| = 0 \quad \text{on } \Gamma. \tag{11.1.15}$$

Here the outward normal derivative $\partial/\partial\nu$ is defined a.e. on Γ, and the boundary conditions (11.1.15) hold at those boundary points where the outward normal vector is defined.

In this example, we obtain a variational inequality from minimizing a non-differentiable energy functional over an entire space. □

Exercise 11.1.1 Show that u is a solution of the constraint minimization problem (11.1.6) if and only if it satisfies the variational inequality (11.1.7).

Exercise 11.1.2 Show that u is a solution of the minimization problem (11.1.12) if and only if it satisfies the variational inequality (11.1.13).

Exercise 11.1.3 Derive the pointwise relations (11.1.14)–(11.1.15) for a smooth solution of the problem (11.1.13). Also show that the boundary conditions (11.1.15) can be expressed as

$$\left|\frac{\partial u}{\partial \nu}\right| \leq g$$

and

$$\left|\frac{\partial u}{\partial \nu}\right| < g \quad \Longrightarrow \quad u = 0,$$

$$\frac{\partial u}{\partial \nu} = g \quad \Longrightarrow \quad u \leq 0,$$

$$\frac{\partial u}{\partial \nu} = -g \quad \Longrightarrow \quad u \geq 0.$$

11.2 Existence and uniqueness based on convex minimization

Convex minimization is a rich source for many elliptic variational inequalities. The discussion in this section relies on materials from Section 3.3 and Section 5.3.

The following result extends Theorem 5.3.19.

Theorem 11.2.1 *Let V be a normed space and $K \subset V$ be a non-empty convex subset. Assume $f : K \to \mathbb{R}$ and $j : K \to \mathbb{R}$ are convex, and f is Gâteaux differentiable. Then*

$$u \in K, \quad f(u) + j(u) = \inf_{v \in K} [f(v) + j(v)] \tag{11.2.1}$$

if and only if

$$u \in K, \quad \langle f'(u), v - u \rangle + j(v) - j(u) \geq 0 \quad \forall\, v \in K. \tag{11.2.2}$$

When K is a subspace and $j(v) \equiv 0$, the inequality (11.2.2) *reduces to an equality:*

$$u \in K, \quad \langle f'(u), v \rangle = 0 \quad \forall\, v \in K. \tag{11.2.3}$$

Proof. Assume u satisfies (11.2.1). Then for any $v \in K$ and any $t \in (0, 1)$, we have $u + t\,(v - u) \in K$ and hence

$$\begin{aligned} f(u) + j(u) &\leq f(u + t\,(v - u)) + j(u + t\,(v - u)) \\ &\leq f(u + t\,(v - u)) + (1 - t)\, j(u) + t\, j(v). \end{aligned}$$

Then,

$$\frac{1}{t} [f(u + t\,(v - u)) - f(u)] + j(v) - j(u) \geq 0 \quad \forall\, t \in (0, 1).$$

Letting $t \to 0+$ we obtain (11.2.2).

Conversely, assume u satisfies (11.2.2). Then since f is convex,

$$f(v) \geq f(u) + \langle f'(u), v - u \rangle.$$

Thus, for any $v \in K$,

$$\begin{aligned} f(v) + j(v) &\geq f(u) + \langle f'(u), v - u \rangle + j(v) \\ &\geq f(u) + j(u). \end{aligned}$$

When K is a subspace and $j(v) \equiv 0$, (11.2.2) is reduced to (11.2.3); see the proof of the same result for Theorem 5.3.19. □

As a corollary of Theorem 3.3.12 and Theorem 11.2.1, we have the next result.

Theorem 11.2.2 *Let V be a Hilbert space, and let $K \subset V$ be non-empty, closed and convex. Assume $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a continuous, symmetric, V-elliptic bilinear form, $\ell \in V'$, and $j : K \to \mathbb{R}$ is convex and l.s.c. on K. Define*

$$E(v) = \frac{1}{2}\, a(v,v) + j(v) - \ell(v).$$

Then the minimization problem

$$u \in K, \quad E(u) = \inf_{v \in K} E(v)$$

has a unique solution. Moreover, $u \in K$ is a solution of the minimization problem if and only if

$$u \in K, \quad a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \forall\, v \in K.$$

We now apply Theorem 11.2.2 to the variational inequalities in Examples 11.1.1 and 11.1.2.

Example 11.2.3 Continuing Example 11.1.1, we notice that the set

$$K = \{v \in H_0^1(\Omega) \mid v \geq \psi \text{ a.e. in } \Omega\}$$

is non-empty, because the function $\max\{0, \psi\}$ belongs to K. It is easy to verify that the set K is closed and convex, the energy functional

$$E(v) = \int_\Omega \left(\frac{1}{2} \left|\nabla v\right|^2 - f\, v \right) dx,$$

is strictly convex, coercive, and continuous on K. Hence, by Theorem 11.2.2, the minimization problem (11.1.6),

$$u \in K, \quad E(u) = \inf_{v \in K} E(v),$$

has a unique solution $u \in K$. This also shows the equivalent variational inequality

$$u \in K, \quad \int_\Omega \nabla u \cdot \nabla (v-u)\, dx \geq \int_\Omega f (v-u)\, dx \quad \forall\, v \in K$$

has a unique solution $u \in K$. □

Example 11.2.4 Continuing Example 11.1.2, we similarly conclude that the variational inequality (11.1.13)

$$u \in V = H^1(\Omega), \int_\Omega \left[\nabla u \cdot \nabla (v-u) + u\,(v-u)\right] dx + g \int_\Gamma \left(|v| - |u|\right) ds$$
$$\geq \int_\Omega f\,(v-u)\, ds \quad \forall\, v \in V$$

has a unique solution. □

Exercise 11.2.1 In this exercise, we consider an extension of Theorem 11.2.2 to the complex case. Let V be a complex Hilbert space, $K \subset V$ be non-empty, closed and convex. Let $a(\cdot,\cdot) : V \times V \to \mathbb{C}$ be continuous, V-elliptic, linear with respect to the first argument, and

$$a(v,u) = \overline{a(u,v)} \quad \forall\, u, v \in V.$$

Let $\ell : V \to \mathbb{R}$ be linear and continuous, and $j : K \to \mathbb{R}$ be convex and l.s.c. on K. Denote

$$E(v) = \frac{1}{2}\, a(v,v) + j(v) - \ell(v).$$

Prove that the minimization problem

$$u \in K, \quad E(u) = \inf_{v \in K} E(v)$$

has a unique solution, and it is characterized by the variational inequality

$$u \in K, \quad \mathrm{Re}\, a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \forall\, v \in K.$$

11.3 Existence and uniqueness results for a family of EVIs

The variational inequalities studied in Section 11.2 are associated with minimization problems. In this section, we consider more general variational inequalities that are not necessarily related to minimization problems. We start with a result that extends Theorem 11.2.2.

Let V be a real Hilbert space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$. We say an operator $A : V \to V$ is strongly monotone if for some constant $c_0 > 0$,

$$(A(u) - A(v), u - v) \geq c_0 \|u - v\|^2 \quad \forall\, u, v \in V. \tag{11.3.1}$$

We say A is Lipschitz continuous if there exists a constant $M > 0$ such that

$$\|A(u) - A(v)\| \leq M\, \|u - v\| \quad \forall\, u, v \in V. \tag{11.3.2}$$

Let K be a set in the space V and let $j : K \to \mathbb{R}$. We will use the same symbol j for both the given functional and its following extension to V:

$$\begin{cases} j(v) & \text{if } v \in K, \\ +\infty & \text{if } v \in V \backslash K. \end{cases}$$

We say $j : V \to \overline{\mathbb{R}} \equiv \mathbb{R} \cup \{\pm\infty\}$ is proper if $j(v) > -\infty$ for all $v \in V$, and $j(v) \not\equiv \infty$. We say $j : V \to \overline{\mathbb{R}}$ is lower semi-continuous (l.s.c.) if

$$v_n \underset{n\to\infty}{\longrightarrow} v \text{ in } V \quad \Longrightarrow \quad j(v) \leq \liminf_{n\to\infty} j(v_n).$$

Obviously, the extension $j : V \to \overline{\mathbb{R}}$ is l.s.c. if and only if $K \subset V$ is closed and $j : K \to \mathbb{R}$ is l.s.c.

Theorem 11.3.1 *Let V be a real Hilbert space, $K \subset V$ be non-empty, closed and convex. Assume $A : V \to V$ is strongly monotone and Lipschitz continuous, $j : K \to \mathbb{R}$ is convex and l.s.c. Then for any $f \in V$, the elliptic variational inequality*

$$u \in K, \quad (A(u), v - u) + j(v) - j(u) \geq (f, v - u) \quad \forall\, v \in K \tag{11.3.3}$$

has a unique solution. Moreover, the solution u depends Lipschitz continuously on f.

Proof. We first prove the solution uniqueness. Assume there are two solutions $u_1, u_2 \in K$. Then there hold

$$\begin{aligned}(A(u_1), u_2 - u_1) + j(u_2) - j(u_1) &\geq (f, u_2 - u_1),\\ (A(u_2), u_1 - u_2) + j(u_1) - j(u_2) &\geq (f, u_1 - u_2).\end{aligned}$$

Adding the two inequalities, we get

$$-\,(A(u_1) - A(u_2), u_1 - u_2) \geq 0.$$

By the strong monotonicity of A, we deduce that $u_1 = u_2$.

We then prove the existence. We convert the VI into an equivalent fixed-point problem. For any $\theta > 0$, the problem (11.3.3) is equivalent to

$$\begin{aligned}u \in K, \quad &(u, v - u) + \theta\, j(v) - \theta\, j(u)\\ &\geq (u, v - u) - \theta\, (A(u), v - u) + \theta\, (f, v - u) \quad \forall\, v \in K.\end{aligned} \tag{11.3.4}$$

Now for any $u \in K$, consider the problem

$$\begin{aligned}w \in K, \quad &(w, v - w) + \theta\, j(v) - \theta\, j(w)\\ &\geq (u, v - w) - \theta\, (A(u), v - w) + \theta\, (f, v - w) \quad \forall\, v \in K.\end{aligned} \tag{11.3.5}$$

This variational inequality is equivalent to the minimization problem

$$w \in K, \quad E(w) = \inf_{v \in K} E(v), \tag{11.3.6}$$

where

$$E(v) = \frac{1}{2}\, \|v\|^2 + \theta\, j(v) - (u, v) + \theta\, (A(u), v) - \theta\, (f, v).$$

By Lemma 11.3.5 below, there exist a continuous linear form ℓ_j on V and a constant $c_j \in \mathbb{R}$ such that

$$j(v) \geq \ell_j(v) + c_j \quad \forall\, v \in V.$$

Applying Theorem 11.2.2, we see that the problem (11.3.6), and hence the problem (11.3.5), has a unique solution $w = P_\theta u$. Obviously a fixed point

of the mapping P_θ is a solution of the problem (11.3.4). We will see that for sufficiently small $\theta > 0$, $P_\theta : K \to K$ is a contraction and hence has a unique fixed-point by the Banach fixed-point theorem (Theorem 5.1.3).

For any $u_1, u_2 \in K$, let $w_1 = P_\theta u_1$ and $w_2 = P_\theta u_2$. Then we have

$$\begin{aligned}
&(w_1, w_2 - w_1) + \theta\, j(w_2) - \theta\, j(w_1) \\
&\qquad \geq (u_1, w_2 - w_1) - \theta\ (A(u_1), w_2 - w_1) + \theta\, (f, w_2 - w_1), \\
&(w_2, w_1 - w_2) + \theta\, j(w_1) - \theta\, j(w_2) \\
&\qquad \geq (u_2, w_1 - w_2) - \theta\ (A(u_2), w_1 - w_2) + \theta\, (f, w_1 - w_2).
\end{aligned}$$

Adding the two inequalities and simplifying, we get

$$\|w_1 - w_2\|^2 \leq (u_1 - u_2 - \theta\, (A(u_1) - A(u_2)), w_1 - w_2)\,,$$

Hence

$$\|w_1 - w_2\| \leq \|u_1 - u_2 - \theta\, (A(u_1) - A(u_2))\|.$$

Now

$$\begin{aligned}
&\|u_1 - u_2 - \theta\, (A(u_1) - A(u_2))\|^2 \\
&\qquad = \|u_1 - u_2\|^2 - 2\,\theta\, (A(u_1) - A(u_2), u_1 - u_2) + \theta^2 \|A(u_1) - A(u_2)\|^2 \\
&\qquad \leq (1 - 2\,c_0\theta + M^2\theta^2)\, \|u_1 - u_2\|^2.
\end{aligned}$$

Therefore,

$$\|w_1 - w_2\| \leq (1 - 2\,c_0\theta + M^2\theta^2)^{1/2} \|u_1 - u_2\|$$

and so for $\theta \in (0, 2\,c_0/M^2)$, the mapping P_θ is a contraction on the closed set K.

Finally, let $f_1, f_2 \in V$, and denote by u_1, u_2 the corresponding solutions of the variational inequality (11.3.3). Then

$$\begin{aligned}
(A(u_1), u_2 - u_1) + j(u_2) - j(u_1) &\geq (f_1, u_2 - u_1), \\
(A(u_2), u_1 - u_2) + j(u_1) - j(u_2) &\geq (f_2, u_1 - u_2).
\end{aligned}$$

Add the two inequalities,

$$(A(u_1) - A(u_2), u_1 - u_2) \leq (f_1 - f_2, u_1 - u_2).$$

Then

$$\|u_1 - u_2\| \leq \frac{M}{c_0}\, \|f_1 - f_2\|,$$

i.e. the solution u depends Lipschitz continuously on f. □

Remark 11.3.2 From the proof, we see that the assumptions on the operator A can be weakened to strong monotonicity and Lipschitz continuity over the set K, i.e., we only need to require the two inequalities (11.3.1) and (11.3.2) for any $u, v \in K$. In related applications, these inequalities on A are usually valid over the entire space V. Moreover, when they are valid only on K, there is usually a natural extension of A on K to an operator A_0 on V such that A_0 is strongly monotone and Lipschitz continuous over the entire space V.

Remark 11.3.3 In Theorem 11.3.1, A is assumed to be Lipschitz continuous. In fact, this condition can be weakened to a local Lipschitz continuity condition for the solution existence and uniqueness of the variational inequality (11.3.3). See Exercise 11.3.1.

Remark 11.3.4 By the Riesz representation theorem, there is a bijection between $\ell \in V'$ and $f \in V$ through the equality $\ell(v) = (f, v)\ \forall\, v \in V$. So it is equally well to express the right hand side of the inequality (11.3.3) by $\ell(v - u)$ for some $\ell \in V'$.

In the proof of Theorem 11.3.9, we applied the following result.

Lemma 11.3.5 *Let V be a normed space. Assume $j : V \to \overline{\mathbb{R}}$ is proper, convex and l.s.c. Then there exist a continuous linear functional $\ell_j \in V'$ and a constant $c_j \in \mathbb{R}$ such that*

$$j(v) \geq \ell_j(v) + c_j \quad \forall\, v \in V.$$

Proof. Since $j : V \to \overline{\mathbb{R}}$ is proper, there exists $v_0 \in V$ with $j(v_0) \in \mathbb{R}$. Choose a real number $a_0 < j(v_0)$. Consider the set

$$J = \{(v, a) \in V \times \mathbb{R} \mid j(v) \leq a\}.$$

This set is called the *epigraph* of j in the literature on convex analysis. From the convexity of j, it is easy to verify that J is closed in $V \times \mathbb{R}$. Since j is l.s.c., it is readily seen that J is closed in $V \times \mathbb{R}$. Thus the sets $\{(v_0, a_0)\}$ and J are disjoint, $\{(v_0, a_0)\}$ is a convex compact set, and J is a closed convex set. By Theorem 3.3.7, we can strictly separate the sets $\{(v_0, a_0)\}$ and J: There exists a non-zero continuous linear functional ℓ on V and $\alpha \in \mathbb{R}$ such that

$$\ell(v_0) + \alpha\, a_0 < \ell(v) + \alpha\, a \quad \forall\, (v, a) \in J. \tag{11.3.7}$$

Taking $v = v_0$ and $a = j(v_0)$ in (11.3.7), we obtain

$$\alpha\, (j(v_0) - a_0) > 0.$$

Thus $\alpha > 0$. Divide the inequality (11.3.7) by α and let $a = j(v)$ to obtain

$$j(v) > \frac{-1}{\alpha}\,\ell(v) + a_0 + \frac{1}{\alpha}\,\ell(v_0).$$

Hence the result follows with $\ell_j(v) = -\ell(v)/\alpha$ and $c_j = a_0 + \ell(v_0)/\alpha$. □

We now consider a special case of Theorem 11.3.1 by setting $j(v) \equiv 0$. Then the variational inequality (11.3.3) is reduced to

$$u \in K, \quad (A(u), v - u) \geq (f, v - u) \quad \forall\, v \in K. \tag{11.3.8}$$

In the literature, such a variational inequality is said to be of the first kind. The obstacle problem, Example 11.1.1, is a representative example of an EVI of the first kind. A variational inequality of the first kind is characterized by the feature that it is posed over a convex subset. When the set K is in fact a subspace of V, the variational inequality becomes a variational equation.

As a corollary of Theorem 11.3.1, we have the following result for the unique solvability of the variational inequality (11.3.8).

Theorem 11.3.6 *Let V be a real Hilbert space, and $K \subset V$ be non-empty, closed and convex. Assume $A : V \to V$ is strongly monotone and Lipschitz continuous. Then for any $f \in V$, the variational inequality* (11.3.8) *has a unique solution $u \in K$, which depends Lipschitz continuously on f.*

Theorem 11.3.6 is a generalization of the Lax-Milgram Lemma (Theorem 8.3.4) for the unique solvability of a linear elliptic boundary value problem; proof of this statement is left as Exercise 11.3.2.

Theorem 11.3.1 can be stated in a form without the explicit use of the set K. For this purpose, we use the extension of the functional j from K to V. Then we can rewrite the variational inequality (11.3.12) as

$$u \in V, \quad (A(u), v - u) + j(v) - j(u) \geq (f, v - u) \quad \forall\, v \in V. \tag{11.3.9}$$

Such a variational inequality is said to be of the second kind, that is featured by the presence of a non-differentiable term in the formulation. Example 11.1.2 provides an example of an EVI of the second kind. Note that without the presence of the nondifferentiable term $j(\cdot)$, the inequality (11.3.9) reduces to an equation.

From Theorem 11.3.1, we obtain the following result.

Theorem 11.3.7 *Let V be a real Hilbert space. Assume $A : V \to V$ is strongly monotone and Lipschitz continuous, and $j : V \to \overline{\mathbb{R}}$ is a proper, convex and l.s.c. functional on V. Then for any $f \in V$, the EVI of the second kind* (11.3.9) *has a unique solution.*

Now we introduce an interesting result for variational inequalities, called Minty's lemma ([174]). The result is useful in proving existence of solutions for some variational inequalities ([144, Chapter III]). As we will see in the next section, Minty's lemma is also useful in proving convergence of numerical solutions for variational inequalities. The next result is Minty's lemma for (11.3.3).

Lemma 11.3.8 (Minty Lemma) *Assume the conditions stated in Theorem 11.3.1; the Lipschitz continuity of A can be weakened to continuity on finite dimensional spaces. Then u is a solution of the variational inequality* (11.3.3) *if and only if*

$$u \in K, \quad (A(v), v-u) + j(v) - j(u) \geq (f, v-u) \quad \forall\, v \in K. \tag{11.3.10}$$

Proof. Let u satisfy (11.3.3). Using the monotonicity of A, we have

$$(A(v), v-u) \geq (A(u), v-u) \quad \forall\, v \in K.$$

Hence, u satisfies (11.3.10).

Conversely, let (11.3.10) hold. For any $v \in K$ and $t \in (0,1)$, we have $u + t\,(v-u) \in K$. Replacing v by this element in (11.3.10), we have

$$t\,(A(u+t(v-u)), v-u) + j(u+t(v-u)) - j(u) \geq t\,\ell(v-u). \tag{11.3.11}$$

By the convexity of j,

$$j(u+t(v-u)) \leq t\,j(v) + (1-t)\,j(u).$$

Use this in (11.3.11), divide by t, and let $t \to 0+$ to obtain (11.3.3). □

In many applications, the operator A is linear and corresponds to a bilinear form on V:

$$(A(u), v) = a(u, v), \quad u, v \in V.$$

In this case, it is convenient to replace (f, v) by $\ell(v)$ for $\ell \in V'$, the space of linear continuous functionals on V. As a direct consequence of Theorem 11.3.1, we have the next result.

Theorem 11.3.9 *Let V be a real Hilbert space, and $K \subset V$ be non-empty, closed and convex. Assume $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a continuous, V-elliptic bilinear form, $\ell \in V'$, and $j : K \to \mathbb{R}$ is convex and l.s.c. Then each of the following elliptic variational inequalities*

$$u \in K, \quad a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \forall\, v \in K, \tag{11.3.12}$$

$$u \in K, \quad a(u, v-u) \geq \ell(v-u) \quad \forall\, v \in K, \tag{11.3.13}$$

$$u \in V, \quad a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \forall\, v \in V \tag{11.3.14}$$

has a unique solution. Moreover, the solution u depends Lipschitz continuously on ℓ.

Example 11.3.10 The obstacle problem (11.1.7) is clearly an elliptic variational inequality of the first kind. By Theorem 11.3.9, the obstacle problem (11.1.7) has a unique solution. □

Example 11.3.11 Applying Theorem 11.3.9, we conclude that the variational inequality considered in Example 11.1.2 has a unique solution. □

We remark that in the case when the bilinear form is symmetric, the variational inequality (11.3.12) is equivalent to the constrained minimization problem

$$\inf_{v\in K}\left[\frac{1}{2}\,a(v,v)+j(v)-\ell(v)\right].$$

This problem has a unique solution by Theorem 11.2.2. Furthermore, we have a useful characterization of the solution of the variational inequality (11.3.13). Let $w\in V$ be the unique solution of the linear elliptic boundary value problem

$$w\in V,\quad a(w,v)=\ell(v)\quad \forall\, v\in V.$$

Then for any $v\in K$, we have

$$a(u,v-u)\geq \ell(v-u)=a(w,v-u),$$

i.e.,

$$a(w-u,v-u)\leq 0\quad \forall\, v\in V.$$

Hence, $u\in K$ is the unique best approximation in K of $w\in V$ with respect to the inner product $(\cdot,\cdot)_a$:

$$\|w-u\|_a=\inf_{v\in K}\|w-v\|_a,$$

where $\|\cdot\|_a=a(\cdot,\cdot)^{1/2}$. This result suggests a possible approach to solve the variational inequality (11.3.13). In the first step, we solve a corresponding linear boundary value problem to get a solution w. In the second step, we compute the projection of w onto K, with respect to the inner product $a(\cdot,\cdot)$.

Solution regularity plays an important role for convergence order of numerical solutions. It is more difficult to study solution regularity for variational inequalities than for ordinary boundary value problems of partial differential equations. In general, regularity of solutions of variational inequalities is limited no matter how smooth are the data. Concerning the solution regularity of the obstacle problem, the following result holds.

Theorem 11.3.12 *Let $\Omega\subset\mathbb{R}^d$ be a $C^{1,1}$ domain, $a_{ij}\in C(\overline{\Omega})$ satisfy*

$$\sum_{i,j=1}^{d}a_{ij}(\boldsymbol{x})\,\xi_i\xi_j\geq\alpha\,|\boldsymbol{\xi}|^2\quad\forall\,\boldsymbol{x}\in\Omega,\,\forall\,\boldsymbol{\xi}\in\mathbb{R}^d$$

for some constant $\alpha > 0$, and for some $p \in [2, \infty)$, $f \in L^p(\Omega)$, $\psi \in W^{2,p}(\Omega)$ with $\psi \leq 0$ on Γ. Denote

$$K = \left\{ v \in H_0^1(\Omega) \mid v \geq \psi \text{ a.e. in } \Omega \right\}.$$

Then the solution of the problem

$$u \in K, \quad \int_\Omega \sum_{i,j=1}^{d} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial (v-u)}{\partial x_j} \, dx \geq \int_\Omega f \, (v-u) \, dx \quad \forall \, v \in K$$

has the regularity $u \in W^{2,p}(\Omega)$; moreover, there exists a constant C_p independent of u, f and ψ such that

$$\|u\|_{2,p} \leq C_p \left(\|f\|_{0,p} + \|\psi\|_{2,p} \right).$$

A proof of this result can be found in [50, Section 3.3], that uses a solution regularity result for a linear elliptic boundary value problem (see [89, Theorem 9.14]).

We now consider an example that indicates the solution regularity stated in Theorem 11.3.12 for the obstacle problem cannot be improved. For this purpose, consider a one-dimensional obstacle problem with $\Omega = (-1, 1)$, $\psi(x) = 1 - 4x^2$, and $f(x) = 0$. Because of the symmetry of the problem, we only need to consider the solution for $x \in [0, 1]$. By Theorem 11.3.12, the solution $u \in W^{2,p}(\Omega)$ for any $p < \infty$. Thus $u \in C^1(\overline{\Omega})$. So the conditions (11.1.10) hold and we have

$$u - \psi \geq 0, \quad -u'' \geq 0, \quad (u - \psi)(-u'') = 0 \quad \text{a.e. in } (-1, 1). \tag{11.3.15}$$

Since ψ is strictly concave, from (11.3.15), we derive the solution formula

$$u(x) = \begin{cases} 1 - 4x^2, & 0 \leq x \leq x_0, \\ \dfrac{1 - 4x_0^2}{x_0 - 1} (x - 1), & x_0 \leq x \leq 1 \end{cases}$$

for some $x_0 \in (0, 1)$. From the continuity of u' at x_0, we can find $x_0 = 1 - \sqrt{3}/2$. Easily,

$$u''(x) = \begin{cases} -8, & 0 < x < x_0, \\ 0, & x_0 < x < 1. \end{cases}$$

We see that u'' has a jump discontinuity at $x = x_0$, and hence, u is not three times weakly differentiable. Because of the limited degree of solution smoothness, usually it is not advisable to use high order numerical methods to solve variational inequalities.

Comments on a history of the regularity theory for the obstacle problems are found in [50, page 73]. In this chapter, our focus is on the existence, uniqueness and numerical approximations.

Exercise 11.3.1 In this exercise, we extend Theorem 11.3.1 to the case when the Lipschitz condition (11.3.2) is replaced by a local Lipschitz condition:

$$\|A(v) - A(w)\| \le m(r)\,\|v - w\| \quad \forall\, v, w \in V_r \equiv \{v \in V \mid \|v\| \le r\},$$

where $r > 0$ is arbitrary and $m : \mathbb{R}_+ \to \mathbb{R}_+$ is non-decreasing. The symbol $\mathbb{R}_+$ stands for the set of positive numbers.
(a) Show that for some non-decreasing function $\tilde{m} : \mathbb{R}_+ \to \mathbb{R}_+$,

$$|(A(v), w)| \le \tilde{m}(\|v\|)\,\|w\| \quad \forall\, v, w \in V.$$

(b) Show that for $r > 0$ large enough, $K_r = K \cap V_r \neq \emptyset$, and the variational inequality

$$u_r \in K_r, \quad (A(u_r), v - u_r) + j(v) - j(u_r) \ge (f, v - u_r) \quad \forall\, v \in K_r$$

has a unique solution.
(c) Show that there exists a number $R_0 > 0$ such that $\|u_r\| < R_0$ for any r sufficiently large.
(d) Denote $u_0 = u_{R_0}$, $K_0 = K_{R_0}$. Then

$$u_0 \in K_0, \quad (A(u_0), v - u_0) + j(v) - j(u_0) \ge (f, v - u_0) \quad \forall\, v \in K_0. \quad (11.3.16)$$

Show that the inequality in (11.3.16) is valid for any $v \in K$.
Hint: For any fixed $v \in K$ with $\|v\| > R_0$, $v_\lambda = (1 - \lambda)\,u_0 + \lambda\, v \in K$ for $\lambda \in (0, 1)$. Take $\lambda = (R_0 - \|u_0\|)/(\|v\| - \|u_0\|) \in (0, 1)$; then $v_\lambda \in K_0$. Use this v_λ in (11.3.16).

Exercise 11.3.2 Show that Theorem 11.3.6 generalizes the Lax-Milgram Lemma.

Exercise 11.3.3 A subset $K \subset V$ is said to be a cone in V if for any $v \in K$ and any $\alpha > 0$, we have $\alpha\, v \in K$. Show that if K is a closed convex cone of V, then the variational inequality (11.3.13) is equivalent to the relations

$$\begin{aligned} &a(u, v) \ge \ell(v) \quad \forall\, v \in K, \\ &a(u, u) = \ell(u). \end{aligned}$$

Exercise 11.3.4 Consider the one-dimensional variational inequality

$$u \in K, \quad \int_0^1 u'(v - u)'dx \ge \int_0^1 g\,(v - u)\,dx \quad \forall\, v \in K,$$

where $g \in L^1(0, 1)$ is a given function, and the set K is defined by

$$K = \{v \in H^1(0, 1) \mid v(0) = 0,\ v(1) \le 0\}.$$

(a) Verify that there exists a unique solution.
(b) Show that the classical formulation is (assuming g is smooth)

$$\begin{aligned} &-u'' = g \quad \text{in } (0, 1), \\ &u(0) = 0,\ u(1) \le 0,\ u'(1) \le 0,\ u(1)\,u'(1) = 0. \end{aligned}$$

Hint: First perform an integration by parts on the left side integral.
(c) In the case where g is a constant, derive the following solution formula

$$u(x) = \begin{cases} -\frac{g}{2}\,x^2 + \frac{g}{2}\,x & \text{if } g \geq 0, \\ -\frac{g}{2}\,x^2 + g\,x & \text{if } g < 0. \end{cases}$$

(d) Find a solution formula for a general $g \in L^1(0,1)$.

Exercise 11.3.5 Consider the one-dimensional variational inequality

$$u \in K, \quad \int_0^1 \left[u'(v-u)' + u\,(v-u)\right] dx \geq \int_0^1 g\,(v-u)\,dx \quad \forall\, v \in K,$$

where $g \in L^1(0,1)$ is a given function, and the set K is defined by

$$K = \{v \in H^1(0,1) \mid v(1) \leq 0\}.$$

(a) Verify that there exists a unique solution.
(b) Show the classical formulation is (assuming g is smooth)

$$\begin{aligned} &-u'' + u = g \quad \text{in } (0,1), \\ &\quad u'(0) = 0,\ u(1) \leq 0,\ u'(1) \leq 0,\ u(1)\,u'(1) = 0. \end{aligned}$$

(c) In the case where g is a constant, derive the following solution formula:

$$u(x) = \begin{cases} g - g\,\dfrac{e^x + e^{-x}}{e + e^{-1}} & \text{if } g \geq 0, \\ g & \text{if } g < 0. \end{cases}$$

Exercise 11.3.6 As another example of EVI of the first kind, we consider the elasto-plastic torsion problem. Let $\Omega \subset \mathbb{R}^2$ be a domain with a Lipschitz continuous boundary Γ. Let

$$\begin{aligned} V &= H_0^1(\Omega), \\ a(u,v) &= \int_\Omega \nabla u \cdot \nabla v\, dx, \\ \ell(v) &= \int_\Omega f\, v\, dx, \\ K &= \{v \in V \mid |\nabla v| \leq 1 \text{ a.e. in } \Omega\}. \end{aligned}$$

Then the problem is

$$u \in K, \quad a(u, v-u) \geq \ell(v-u) \quad \forall\, v \in K.$$

Use Theorem 11.3.6 to show that the elasto-plastic torsion problem has a unique solution. When Ω is a smooth domain, it can be shown that the variational inequality problem is equivalent to one over the set

$$K = \{v \in V \mid |v(\boldsymbol{x})| \leq \operatorname{dist}(\boldsymbol{x}, \Gamma) \text{ a.e. in } \Omega\}$$

with the same bilinear and linear forms. Here, $\text{dist}(\boldsymbol{x}, \Gamma)$ is the distance between $\boldsymbol{x}$ and Γ.

In general one cannot expect high regularity for the solution of the elasto-plastic torsion problem. It can be shown that if Ω is convex or smooth and $f \in L^p(\Omega)$, $1 < p < \infty$, then the solution of the elasto-plastic torsion problem $u \in W^{2,p}(\Omega)$. The following exact solution shows that $u \notin H^3(\Omega)$ even if Ω and f are smooth ([91, Chapter 2]).

Consider the special situation where Ω is the circle centered at O with the radius R, and the external force density $f = c_0 > 0$ is a constant. Denote $r = \|\boldsymbol{x}\|_2$. Verify that if $c_0 \le 2/R$, then the solution is given by

$$u(\boldsymbol{x}) = (c_0/4)\,(R^2 - r^2);$$

where if $c_0 > 2/R$, then

$$u(\boldsymbol{x}) = \begin{cases} (c_0/4)\,[R^2 - r^2 - (R - 2/c_0)^2], & \text{if } 0 \le r \le 2/c_0, \\ R - r, & \text{if } 2/c_0 \le r \le R. \end{cases}$$

Exercise 11.3.7 A simplified version of the Signorini problem can be described as an EVI of the first kind with the following data

$$\begin{aligned} V &= H^1(\Omega), \\ K &= \{v \in V \mid v \ge 0 \text{ a.e. on } \Gamma\}, \\ a(u, v) &= \int_\Omega (\nabla u \cdot \nabla v + u\, v)\, dx, \\ \ell(v) &= \int_\Omega f\, v\, dx + \int_\Gamma g\, v\, ds, \end{aligned}$$

where $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$ are given functions. Show that the corresponding variational inequality problem

$$u \in K, \quad a(u, v - u) \ge \ell(v - u) \quad \forall\, v \in K$$

has a unique solution u. Show that formally, u solves the boundary value problem

$$\begin{gathered} -\Delta u + u = f \quad \text{a.e. in } \Omega, \\ u \ge 0, \quad \frac{\partial u}{\partial \nu} \ge g, \quad u\left(\frac{\partial u}{\partial \nu} - g\right) = 0 \quad \text{a.e. on } \Gamma. \end{gathered}$$

Exercise 11.3.8 The double obstacle problem is

$$u \in K_{\phi,\psi}, \quad a(u, v - u) \ge \ell(v - u) \quad \forall\, v \in K_{\phi,\psi},$$

where

$$K_{\phi,\psi} = \left\{v \in H_0^1(\Omega) \mid \phi(\boldsymbol{x}) \le v(\boldsymbol{x}) \le \psi(\boldsymbol{x}) \text{ a.e. in } \Omega\right\}$$

with two given measurable functions ϕ and ψ. Assume $K_{\phi,\psi} \ne \emptyset$, $a(\cdot, \cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is bilinear, continuous and $H_0^1(\Omega)$-elliptic, $\ell \in H^{-1}(\Omega)$. Show that the double obstacle problem has a unique solution.

Exercise 11.3.9 Consider the equilibrium position of two membranes subject to applied transversal forces of density f_1 and f_2. The set is

$$K = \{\boldsymbol{v} = (v_1, v_2)^T \in H^1(\Omega)^2 \mid v_1 \le v_2 \text{ a.e. in } \Omega,\ v_i = \phi_i \text{ on } \Gamma,\ i = 1, 2\}.$$

Here, ϕ_1, ϕ_2, c_1, c_2, f_1 and f_2 are given, $\phi_1 \le \phi_2$ in $\overline{\Omega}$. The constraint "$v_1 \le v_2$ in Ω" expresses the assumption that the two membranes do not penetrate each other. The variational inequality of the problem is

$$\begin{aligned} \boldsymbol{u} \in K, \quad & \sum_{i=1}^{2} \int_\Omega \left[\nabla u_i \cdot \nabla(v_i - u_i) + c_i u_i (v_i - u_i)\right] dx \\ & \ge \sum_{i=1}^{2} \int_\Omega f_i (v_i - u_i)\, dx \quad \forall\, \boldsymbol{v} = (v_1, v_2)^T \in K. \end{aligned}$$

Provide conditions on the data that guarantee the unique solvability of the variational inequality. Show that the variational inequality is equivalent to the minimization problem

$$\boldsymbol{u} \in K, \quad E(\boldsymbol{u}) = \inf\{E(\boldsymbol{v}) \mid \boldsymbol{v} \in K\},$$

where

$$E(\boldsymbol{v}) = \frac{1}{2} \sum_{i=1}^{2} \int_\Omega \left[|\nabla v_i|^2 + c_i (v_i)^2\right] dx - \sum_{i=1}^{2} \int_\Omega f_i v_i dx.$$

Exercise 11.3.10 In addition to the conditions stated in Theorem 11.3.7, assume $j : V \to \mathbb{R}$ is positively homogeneous. Show that u is the solution of the variational inequality (11.3.14) if and only if it satisfies the two relations:

$$\begin{aligned} & a(u, v) + j(v) \ge \ell(v) \quad \forall\, v \in V, \\ & a(u, u) + j(u) = \ell(u). \end{aligned}$$

Exercise 11.3.11 Consider the one-dimensional variational inequality

$$u \in V, \quad \int_0^1 u'(v - u)' dx + g\,[|v(1)| - |u(1)|] \ge \int_0^1 f\,(v - u)\, dx \quad \forall\, v \in V,$$

where $g > 0$ is a given constant, $f \in L^1(0, 1)$ is a given function, and the space V is defined by

$$V = \{v \in H^1(0, 1) \mid v(0) = 0\}.$$

(a) Show that there is a unique solution.
(b) Derive the classical formulation (assuming f is smooth):

$$\begin{aligned} & -u'' = f \quad \text{in } (0, 1), \\ & u(0) = 0,\ |u'(1)| \le g,\ u'(1)\, u(1) + g\,|u(1)| = 0. \end{aligned}$$

(c) In the case f is a constant, show the solution formula

$$u(x) = \begin{cases} -\dfrac{f}{2}\, x^2 + (f - g)\, x & \text{if } f \ge 2\, g, \\ -\dfrac{f}{2}\, x^2 + \dfrac{f}{2}\, x & \text{if } |f| < 2\, g, \\ -\dfrac{f}{2}\, x^2 + (f + g)\, x & \text{if } f \le -2\, g. \end{cases}$$

(d) Find a solution formula for a general $f \in L^1(0,1)$.

Exercise 11.3.12 The problem of the flow of a viscous plastic fluid in a pipe can be formulated as an EVI of the second kind. Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain. Define

$$\begin{aligned} V &= H_0^1(\Omega), \\ a(u,v) &= \mu \int_\Omega \nabla u \cdot \nabla v \, dx, \\ \ell(v) &= \int_\Omega f\, v \, dx, \\ j(v) &= g \int_\Omega |\nabla v| \, dx, \end{aligned}$$

where $\mu > 0$ and $g > 0$ are two parameters. Then the problem is

$$u \in V, \quad a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \forall\, v \in V.$$

Show that the problem has a unique solution.

Exercise 11.3.13 Assume V is a Hilbert space, $A : V \to V$ is strongly monotone and uniformly Lipschitz continuous. Then for any $f \in V$, the equation $A(u) = f$ has a unique solution $u \in V$, and the solution depends Lipschitz continuously on f.

11.4 Numerical approximations

Consider the numerical approximation of the general EVI (11.3.3), which is rewritten as

$$u \in K, \quad (A(u), v-u) + j(v) - j(u) \geq \ell(v-u) \quad \forall\, v \in K. \tag{11.4.1}$$

Owing to the Riesz representation theorem, Theorem 2.5.8, there is a one-to-one correspondence between $f \in V$ and $\ell \in V'$. As in Theorem 11.3.1, assume V is a real Hilbert space, $K \subset V$ is non-empty, convex and closed, $A : V \to V$ is strongly monotone and Lipschitz continuous, $j : K \to \mathbb{R}$ is convex and l.s.c., and $\ell \in V'$. Then by Theorem 11.3.1, the variational inequality (11.4.1) has a unique solution.

Before considering a general framework for the numerical solution of the variational inequality (11.4.1), we make a further assumption on the functional j:

$$j(v) = j_0(v) \text{ for } v \in K, \quad j_0 : V \to \mathbb{R} \text{ is convex and l.s.c.} \tag{11.4.2}$$

In other words, j is the restriction of j_0 on K, and $j_0 : V \to \mathbb{R}$ is convex and l.s.c. on V. This assumption is usually naturally satisfied. For example,

consider the following variational inequality

$$u \in K, \quad \int_\Omega [\nabla u \cdot \nabla(v-u) + u\,(v-u)]\,dx + g\int_\Gamma (|v| - |u|)\,ds \geq \int_\Omega f\,(v-u)\,dx \quad \forall\, v \in K,$$

where $K = \{v \in V \mid v \geq \psi \text{ in } \Omega\}$, $V = H^1(\Omega)$, and $\psi \in V$ is a given function. Here

$$j(v) = \begin{cases} g\displaystyle\int_\Gamma |v|\,ds, & v \in K, \\ +\infty, & v \notin K. \end{cases}$$

For this example, we may take

$$j_0(v) = g\int_\Gamma |v|\,ds, \quad v \in V.$$

It is convex and continuous over V.

Let $V_h \subset V$ be a finite element space, and let $K_h \subset V_h$ be non-empty, convex and closed. Then the finite element approximation of the problem (11.4.1) is

$$u_h \in K_h, \quad (A(u_h), v_h - u_h) + j_0(v_h) - j_0(u_h) \geq \ell(v_h - u_h) \quad \forall\, v_h \in K_h. \tag{11.4.3}$$

Another application of Theorem 11.3.1 shows that the discrete problem (11.4.3) has a unique solution under the stated assumptions on the given data.

We have the following general convergence result of the finite element method. It extends some convergence results found in the literature ([91]).

Theorem 11.4.1 *In addition to the assumptions made in Theorem 11.3.1 and* (11.4.2)*, we further assume $j_0 : V \to \mathbb{R}$ is continuous, and $\{K_h\}_h$ approximates the set K in the following sense:*

(a) *For any $v \in K$, there exists $v_h \in K_h$ such that $\|v_h - v\|_V \to 0$ as $h \to 0$.*

(b) *$v_h \in K_h$ with $v_h \rightharpoonup v$ as $h \to 0$ implies $v \in K$.*

Then we have convergence $\|u_h - u\|_V \to 0$ as $h \to 0$.

Proof. The proof is divided into three steps.

In the first step, we show boundedness of the set $\{u_h\}_h$ in V. For this purpose, we fix a $v_0 \in K$ and select $v_{0,h} \in K_h$ such that $v_{0,h} \to v_0$ in V as $h \to 0$. Let $v_h = v_{0,h}$ in (11.4.3) to get

$$\begin{aligned}(A(u_h) - A(v_{0,h}), u_h - v_{0,h}) \leq\ & (A(v_{0,h}) - A(v_0), v_{0,h} - u_h) \\ & + (A(v_0), v_{0,h} - u_h) \\ & + j_0(v_{0,h}) - j_0(u_h) - \ell(v_{0,h} - u_h).\end{aligned}$$

Using the monotonicity and Lipschitz continuity of A, and applying Lemma 11.3.5 on the term $-j_0(u_h)$, we obtain from the above inequality that

$$\begin{aligned} c_0\|u_h - v_{0,h}\|^2 &\\ \le \{M\,[\|A(v_{0,h}) - A(v_0)\| + \|A(v_0)\|] + \|\ell\| + \|\ell_{j_0}\|\}\,\|u_h - v_{0,h}\| &\\ + |j_0(v_{0,h}) - j_0(v_0)| + j_0(v_0) + |c_{j_0}| + \|\ell_{j_0}\|\,(\|v_{0,h} - v_0\| + \|v_0\|)\,. & \end{aligned}$$

As $h \to 0$, since $\|v_{0,h} - v_0\| \to 0$, we know $\|A(v_{0,h}) - A(v_0)\| \to 0$ and $|j_0(v_{0,h}) - j_0(v_0)| \to 0$. So $\{\|u_h - v_{0,h}\|\}_h$ is bounded, and then $\{\|u_h\|\}_h$ is bounded. Thus, for a subsequence of $\{u_h\}_h$, still denoted as $\{u_h\}_h$, and some $w \in V$, we have the weak convergence

$$u_h \rightharpoonup w \quad \text{in } V.$$

By Assumption (b), $w \in K$.

In the second step, we prove the weak limit w is the solution of the problem (11.4.1). From Minty Lemma 11.3.8 (in its discrete form), we know (11.4.3) is equivalent to

$$u_h \in K_h, \quad (A(v_h), v_h - u_h) + j_0(v_h) - j_0(u_h) \ge \ell(v_h - u_h) \quad \forall\, v_h \in K_h. \tag{11.4.4}$$

For any fixed $v \in K$, choose $v_h \in K_h$ with $v_h \to v$ in V as $h \to 0$. Then as $h \to 0$,

$$\begin{aligned} &A(v_h) \to A(v), \quad (A(v_h), v_h - u_h) \to (A(v), v - w)\,, \\ &j_0(v_h) \to j_0(v), \quad \ell(v_h - u_h) \to \ell(v - w). \end{aligned}$$

From the l.s.c. of j_0,

$$j_0(w) \le \liminf_{h \to 0} j_0(u_h).$$

Thus taking the limit $h \to 0$ in (11.4.4) and noting $j_0 = j$ on K, we obtain

$$w \in K, \quad (A(v), v - w) + j(v) - j(w) \ge \ell(v - w) \quad \forall\, v \in K.$$

Applying the Minty Lemma again, we see that w is a solution of the problem (11.4.1). Since the problem (11.4.1) has a unique solution u, we conclude $w = u$.

In the last step, we show the strong convergence of u_h towards u. We choose $\tilde{u}_h \in K_h$ such that $\tilde{u}_h \to u$ in V as $h \to 0$. Using the strong monotonicity of A,

$$\begin{aligned} c_0\|u - u_h\|^2 &\le (A(u) - A(u_h), u - u_h) \\ &= (A(u), u - u_h) - (A(u_h), \tilde{u}_h - u_h) - (A(u_h), u - \tilde{u}_h)\,. \end{aligned} \tag{11.4.5}$$

By (11.4.3),

$$-(A(u_h), \tilde{u}_h - u_h) \le j_0(\tilde{u}_h) - j_0(u_h) - \ell(\tilde{u}_h - u_h).$$

So from (11.4.5) we deduce that

$$c_0\|u-u_h\|^2 \le (A(u),u-u_h)+j_0(\tilde{u}_h)-j_0(u_h) \\ -\ell(\tilde{u}_h-u_h)-(A(u_h),u-\tilde{u}_h)\,. \tag{11.4.6}$$

Now as $h\to 0$,

$$(A(u),u-u_h)\to 0,\quad \ell(\tilde{u}_h-u_h)\to 0,\quad (A(u_h),u-\tilde{u}_h)\to 0.$$

Moreover,

$$\lim_{h\to 0} j_0(\tilde{u}_h)=j_0(u),\quad \limsup_{h\to 0}\,[-j_0(u_h)]\le -j_0(u).$$

Thus from (11.4.6), we have

$$\limsup_{h\to 0} c_0\|u-u_h\|^2\le 0.$$

Therefore, u_h converges strongly to u as $h\to 0$. □

Now we turn to error estimation for the finite element solution u_h. We follow [79] and first give an abstract error analysis.

Theorem 11.4.2 *Let*

$$R(v,w)=(A(u),v-w)+j_0(v)-j_0(w)-\ell(v-w).$$

Then

$$\frac{c_0}{2}\|u-u_h\|^2\le \inf_{v\in K} R(v,u_h)+\inf_{v_h\in K_h}\left[R(v_h,u)+\frac{M^2}{2\,c_0}\,\|u-v_h\|^2\right]. \tag{11.4.7}$$

Proof. By the strong monotonicity of A, for any $v_h\in K_h$ we have

$$(A(u)-A(u_h),u-u_h)\ge c_0\|u-u_h\|^2.$$

Add the inequalities in (11.4.1) and (11.4.3) to the above inequality to obtain

$$c_0\|u-u_h\|^2\le R(v,u_h)+R(v_h,u)+(A(u_h)-A(u),v_h-u)\,. \tag{11.4.8}$$

We bound the last term as follows:

$$(A(u_h)-A(u),v_h-u)\le M\,\|u-u_h\|\,\|u-v_h\| \\ \le \frac{c_0}{2}\,\|u-u_h\|^2+\frac{M^2}{2\,c_0}\,\|u-v_h\|^2.$$

This bound combined with (11.4.8) leads to (11.4.7). □

In the case where A is linear, $a(u,v) = (A(u),v)$ is a bilinear form on V. Then (11.4.1) becomes

$$u \in K, \quad a(u, v-u) + j(v) - j(u) \geq \ell(v-u) \quad \forall\, v \in K,$$

and (11.4.3) becomes

$$u_h \in K_h, \quad a(u_h, v_h - u_h) + j_0(v_h) - j_0(u_h) \geq \ell(v_h - u_h) \quad \forall\, v_h \in K_h.$$

We still have the error bound (11.4.7) with

$$R(v,w) = a(u, v-w) + j_0(v) - j_0(w) - \ell(v-w).$$

The inequality (11.4.7) is a generalization of Céa's lemma to the finite element approximation of elliptic variational inequalities of the first kind. It is easy to see that the inequality (11.4.7) reduces to Céa's lemma in the case of finite element approximation of a variational equation problem.

In the case $K_h \subset K$, we have the so-called *internal approximation* of the elliptic variational inequality. Since now $u_h \in K$, the first term on the right hand side of (11.4.7) vanishes, and the error inequality (11.4.7) reduces to

$$\|u - u_h\| \leq c \inf_{v_h \in K_h} \left[\|u - v_h\| + |R(v_h, u)|^{1/2} \right].$$

Example 11.4.3 Let us apply Theorem 11.4.2 to derive an order error estimate for the approximation of the obstacle problem, Example 11.1.1. Such an estimate was first proved in [79]. We assume $u, \psi \in H^2(\Omega)$ and Ω is a polygon, and use linear elements on a mesh of triangles from a regular family of triangulations. Then the discrete admissible set is

$$K_h = \{v_h \in H_0^1(\Omega) \mid v_h \text{ is piecewise linear,} \\ v_h(\boldsymbol{x}) \geq \psi(\boldsymbol{x}) \text{ for any node } \boldsymbol{x}\}.$$

We see that any function in K_h is a continuous piecewise linear function, vanishing on the boundary and dominating the obstacle function at the interior nodes of the mesh. In general, $K_h \not\subset K$. For any $u \in H^2(\Omega)$ and $v, w \in H_0^1(\Omega)$ we have

$$R(v,w) = \int_\Omega \left[\nabla u \cdot \nabla(v-w) - f\,(v-w)\right] dx = \int_\Omega \left(-\Delta u - f\right)(v-w)\, dx.$$

Thus from the inequality (11.4.7), we have the following error bound

$$\|u - u_h\|_1 \leq c \left\{ \inf_{v_h \in K_h} \left[\|u - v_h\|_1 + \| -\Delta u - f\|_0^{1/2} \|u - v_h\|_0^{1/2} \right] \right. \\ \left. + \| -\Delta u - f\|_0^{1/2} \inf_{v \in K} \|v - u_h\|_0^{1/2} \right\} \tag{11.4.9}$$

Let $\Pi_h u$ be the continuous piecewise linear interpolant of u. Then $\Pi_h u \in K_h$ and

$$\begin{aligned}\inf_{v_h \in K_h} \left[\|u - v_h\|_1 + \| - \Delta u - f\|_0^{1/2} \|u - v_h\|_0^{1/2} \right] \\ \leq \|u - \Pi_h u\|_1 + \| - \Delta u - f\|_0^{1/2} \|u - \Pi_h u\|_0^{1/2} \\ \leq c \left[|u|_2 + \| - \Delta u - f\|_0^{1/2} |u|_2^{1/2} \right] h.\end{aligned}$$

To bound the term $\inf_{v \in K} \|v - u_h\|_0$, we define

$$u^{h,*} = \max\{u_h, \psi\}.$$

Since $u_h, \psi \in H^1(\Omega)$, we have $u^{h,*} \in H^1(\Omega)$. By the definition, certainly $u^{h,*} \geq \psi$. Finally, since $\psi \leq 0$ on Γ, we have $u^{h,*} = 0$ on Γ. Hence, $u^{h,*} \in K$. Let

$$\Omega^* = \{\boldsymbol{x} \in \Omega \mid u_h(\boldsymbol{x}) < \psi(\boldsymbol{x})\}.$$

Then over $\Omega \backslash \Omega^*$, $u^{h,*} = u_h$, and so

$$\inf_{v \in K} \|v - u_h\|_0^2 \leq \|u^{h,*} - u_h\|_0^2 = \int_{\Omega^*} |u_h - \psi|^2 dx.$$

Let $\Pi_h \psi$ be the continuous piecewise linear interpolant of ψ. Since at any node, $u_h \geq \psi = \Pi_h \psi$, we have $u_h \geq \Pi_h \psi$ in Ω. Therefore, over Ω^*,

$$0 < |u_h - \psi| = \psi - u_h \leq \psi - \Pi_h \psi = |\psi - \Pi_h \psi|.$$

Thus,

$$\int_{\Omega^*} |u_h - \psi|^2 dx \leq \int_{\Omega^*} |\psi - \Pi_h \psi|^2 dx \leq \int_{\Omega} |\psi - \Pi_h \psi|^2 dx \leq c\, |\psi|_2^2 h^4,$$

and then

$$\inf_{v \in K} \|v - u_h\|_0^{1/2} \leq c\, |\psi|_2^{1/2} h.$$

From the inequality (11.4.9), we finally get the optimal order error estimate

$$\|u - u_h\|_{H^1(\Omega)} \leq c\, h$$

for some constant $c > 0$ depending only on $|u|_2$, $\|f\|_0$ and $|\psi|_2$. □

Example 11.4.4 Let us apply Theorem 11.4.2 to derive an error estimate for some finite element solution of the variational inequality (11.1.13) of Example 11.1.2:

$$u \in V, \quad a(u, v - u) + j(v) - j(u) \geq \ell(v - u) \quad \forall\, v \in V \tag{11.4.10}$$

where

$$\begin{aligned} K &= V = H^1(\Omega), \\ a(u,v) &= \int_\Omega (\nabla u \cdot \nabla v + u\, v)\, dx, \\ \ell(v) &= \int_\Omega f\, v\, dx, \\ j(v) &= g \int_\Gamma |v|\, ds. \end{aligned}$$

Here $f \in L^2(\Omega)$ and $g > 0$ are given, $\Omega \subset \mathbb{R}^2$ is a polygonal domain. We write $\Gamma = \cup_{i=1}^{i_0} \Gamma_i$, where each Γ_i is a line segment. By Theorem 11.3.7, (11.4.10) has a unique solution. Assume

$$u \in H^2(\Omega), \quad u|_{\Gamma_i} \in H^2(\Gamma_i) \ \ \forall\, i.$$

Let V_h be a piecewise linear finite element space constructed from a regular partition of the domain Ω, and let $u_h \in V_h$ be the finite element solution. By an integration by parts,

$$\begin{aligned} R(v_h, u) = &\int_\Gamma \left[\frac{\partial u}{\partial \nu} (v_h - u) + g\, (|v_h| - |u|) \right] ds \\ &+ \int_\Omega (-\Delta u + u - f)\, (v_h - u)\, dx. \end{aligned}$$

Thus,

$$\begin{aligned} |R(v_h, u)| \le &\left[\left\| \frac{\partial u}{\partial \nu} \right\|_{L^2(\Gamma)} + g \sqrt{\text{meas}\,(\Gamma)} \right] \|v_h - u\|_{L^2(\Gamma)} \\ &+ \| -\Delta u + u - f \|_{L^2(\Omega)} \|v_h - u\|_{L^2(\Omega)}. \end{aligned}$$

Using (11.4.7) we get

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} \le c(u) \inf_{v_h \in V_h} \big[&\|u - v_h\|_{H^1(\Omega)} + \|u - v_h\|_{L^2(\Gamma)}^{1/2} \\ &+ \|u - v_h\|_{L^2(\Omega)}^{1/2} \big]. \end{aligned}$$

Then we have the optimal order error estimate

$$\|u - u_h\|_{H^1(\Omega)} \le c(u)\, h. \qquad \square$$

Let us consider more the numerical approximation of EVIs of the second kind:

$$u \in V, \quad (A(u), v - u) + j(v) - j(u) \ge \ell(v - u) \quad \forall\, v \in V. \qquad (11.4.11)$$

Here V is a real Hilbert space, $A : V \to V$ is strongly monotone and Lipschitz continuous, $j(\cdot) : V \to \mathbb{R}$ is convex and l.s.c., and $\ell \in V'$. Let $V_h \subset V$ be a finite element space. Then the finite element approximation of the problem (11.4.11) is

$$u_h \in V_h, \quad (A(u_h), v_h - u_h) + j(v_h) - j(u_h) \geq \ell(v_h - u_h) \quad \forall\, v_h \in V_h. \tag{11.4.12}$$

Both (11.4.11) and (11.4.12) have a unique solution. A major issue in solving the discrete system (11.4.12) is the treatment of the non-differentiable term. In practice, several approaches can be used, e.g., regularization technique, method of Lagrangian multipliers, method of numerical integration. We will briefly describe the regularization technique and the method of Lagrangian multipliers, and provides a detailed discussion of error analysis for the method of numerical integration.

Regularization technique. The basic idea of the regularization method is to approximate the non-differentiable term $j(\cdot)$ by a family of differentiable ones $j_\varepsilon(\cdot)$, where $\varepsilon > 0$ is a small regularization parameter. Convergence of the method is obtained when $\varepsilon \to 0$. Our presentation of the method is given on the continuous level; the extension of the method to the discrete level is straightforward. For the approximate solution of the variational inequality (11.4.11), we introduce the regularized problem

$$u_\varepsilon \in V, \quad (A(u_\varepsilon), v - u_\varepsilon) + j_\varepsilon(v) - j_\varepsilon(u_\varepsilon) \geq \ell(v - u_\varepsilon) \quad \forall\, v \in V. \tag{11.4.13}$$

Since $j_\varepsilon(\cdot)$ is differentiable, the variational inequality (11.4.13) is actually a nonlinear equation:

$$u_\varepsilon \in V, \quad (A(u_\varepsilon), v) + \langle j_\varepsilon'(u_\varepsilon), v\rangle = \ell(v) \quad \forall\, v \in V. \tag{11.4.14}$$

Many possible regularization functions can be used for this purpose. For

$$j(v) = g \int_\Gamma |v|\, ds,$$

we let

$$j_\varepsilon(v) = g \int_\Gamma \phi_\varepsilon(v)\, ds$$

where $\phi_\varepsilon(t)$ is differentiable with respect to t and approximates $|t|$ as $\varepsilon \to 0$. We may choose

$$\phi_\varepsilon(t) = \begin{cases} t - \varepsilon/2 & \text{if } t \geq \varepsilon, \\ t^2/(2\varepsilon) & \text{if } |t| \leq \varepsilon, \\ -t - \varepsilon/2 & \text{if } t \leq -\varepsilon, \end{cases} \tag{11.4.15}$$

or

$$\phi_\varepsilon(t) = \begin{cases} t & \text{if } t \geq \varepsilon, \\ (t^2/\varepsilon + \varepsilon)/2 & \text{if } |t| \leq \varepsilon, \\ -t & \text{if } t \leq -\varepsilon, \end{cases} \tag{11.4.16}$$

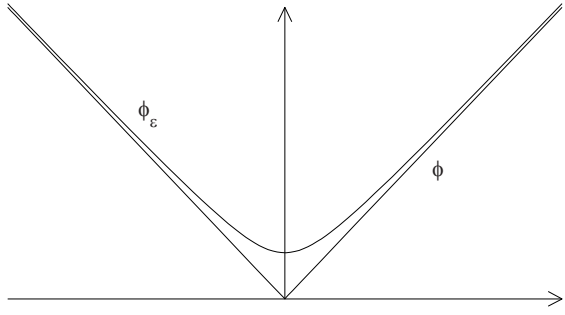

FIGURE 11.1. Regularization function

or

$$\phi_\varepsilon(t) = \sqrt{t^2 + \varepsilon^2}, \tag{11.4.17}$$

or

$$\phi_\varepsilon(t) = \frac{\varepsilon}{\varepsilon + 1} \left(\frac{|t|}{\varepsilon} \right)^{\varepsilon+1}, \tag{11.4.18}$$

or

$$\phi_\varepsilon(t) = \frac{t^{\varepsilon+1}}{\varepsilon + 1} \tag{11.4.19}$$

and the list can be further expanded. Figure 11.1 shows graphs of the functions $\phi(t) = |t|$ and $\phi_\varepsilon(t) = \sqrt{t^2 + \varepsilon^2}$.

A general convergence result for the regularization method can be found in [93], [91]. The regularization method has been widely used in solving variational inequalities involving non-differentiable terms, see, e.g., [143], [193].

It is not difficult to derive a priori error estimates of the form

$$\|u - u_\varepsilon\|_V \leq c\, \varepsilon^\beta \tag{11.4.20}$$

for some exponent $\beta > 0$ (see the references mentioned above). The major problem associated with the regularization method is that the conditioning of a regularized problem deteriorates as $\varepsilon \to 0$. Thus, there is a tradeoff in the selection of the regularization parameter. Theoretically, to get more accurate approximations, we need to use smaller ε. Yet, if ε is too small, the numerical solution of the regularized problem cannot be computed accurately. It is highly desirable to have a posteriori error estimates which can give us computable error bounds once we have solutions of regularized

problems. We can use the a posteriori error estimates in devising a stopping criterion in actual computations: If the estimated error is within the given error tolerance, we accept the solution of the regularized problem as the exact solution; and if the estimated error is large, then we need to use a smaller value for the regularization parameter ε. An adaptive algorithm can be developed based on the a posteriori error analysis. A posteriori error estimates of the form

$$\|u - u_\varepsilon\|_V \leq F(u_\varepsilon),$$

where the error bound can be easily computed once the regularization solution u_ε is known, have been derived in several papers, see, e.g., [109, 112, 114, 127]. For some choices of the regularization function, the constant c and power β can be determined explicitly (Exercise 11.4.3), and then (11.4.20) can serve as an a posteriori error estimate.

Method of Lagrangian multipliers. Again, here our presentation of the method is given on the continuous level. We take the simplified friction problem as an example. Let

$$\Lambda = \{\mu \in L^\infty(\Gamma) \mid |\mu| \leq 1 \text{ a.e. on } \Gamma\}.$$

Following [91], we have the following result.

Theorem 11.4.5 *The simplified friction problem* (11.4.10) *is equivalent to the problem of finding* $u \in V$ *and* $\lambda \in \Lambda$ *such that*

$$\int_\Omega (\nabla u \cdot \nabla v + u\, v)\, dx + g \int_\Gamma \lambda\, v\, ds = \int_\Omega f\, v\, dx \quad \forall\, v \in V, \tag{11.4.21}$$

$$\lambda\, u = |u| \quad \text{a.e. on } \Gamma. \tag{11.4.22}$$

λ *is called a Lagrangian multiplier, and is unique.*

Proof. Let u be the solution of the variational inequality (11.4.10). Then from Exercise 11.3.10 we have

$$a(u, u) + j(u) = \ell(u) \tag{11.4.23}$$

and

$$a(u, v) + j(v) \geq \ell(v) \quad \forall\, v \in V.$$

The latter relation implies

$$|\ell(v) - a(u, v)| \leq j(v) \quad \forall\, v \in V. \tag{11.4.24}$$

Denote $L(v) = \ell(v) - a(u, v)$. Then the value of $L(v)$ depends on the trace $v|_\Gamma$ only and $L(\cdot)$ is a linear continuous functional on $H^{1/2}(\Gamma)$. Moreover, we obtain from (11.4.24) the estimate

$$|L(v)| \leq g\, \|v\|_{L^1(\Gamma)} \quad \forall\, v \in H^{1/2}(\Gamma).$$

Since $H^{1/2}(\Gamma)$ is a subspace of $L^1(\Gamma)$, applying the Hahn-Banach theorem we can extend the functional L to $\tilde{L} \in (L^1(\Gamma))'$ such that

$$\|\tilde{L}\| = \|L\| \le g.$$

Since $(L^1(\Gamma))' = L^\infty(\Gamma)$, we have the existence of a $\lambda \in \Lambda$ such that

$$\tilde{L}(v) = g \int_\Gamma \lambda\, v\, ds \quad \forall\, v \in L^1(\Gamma).$$

Therefore,

$$\ell(v) - a(u,v) = L(v) = \tilde{L}(v) = g \int_\Gamma \lambda\, v\, ds \quad \forall\, v \in V,$$

i.e. (11.4.21) holds.

Taking $v = u$ in (11.4.21) we obtain

$$a(u,u) + g \int_\Gamma \lambda\, u\, ds = \ell(u).$$

This relation and (11.4.23) together imply

$$\int_\Gamma (|u| - \lambda\, u)\, ds = 0.$$

Since $|\lambda| \le 1$ a.e. on Γ, we must have (11.4.22).

For the uniqueness of λ, suppose (11.4.21) holds with λ replaced by $\tilde{\lambda} \in \Lambda$. Then

$$\int_\Gamma (\lambda - \tilde{\lambda})\, v\, ds = 0 \quad \forall\, v \in V.$$

By the density of the trace space $\gamma(V) = H^{1/2}(\Gamma)$ in $L^2(\Gamma)$, we obtain

$$\int_\Gamma (\lambda - \tilde{\lambda})\, v\, ds = 0 \quad \forall\, v \in L^2(\Gamma).$$

Then take $v = \lambda - \tilde{\lambda}$ to conclude $\tilde{\lambda} = \lambda$.

Conversely, suppose we have $u \in V$ and $\lambda \in \Lambda$ satisfying (11.4.21) and (11.4.22). Then using (11.4.21) with v replaced by $v - u$, we obtain

$$a(u, v-u) + g \int_\Gamma \lambda\, v\, ds - g \int_\Gamma \lambda\, u\, ds = \ell(v-u).$$

Noticing that

$$g \int_\Gamma \lambda\, u\, ds = g \int_\Gamma |u|\, ds = j(u),$$
$$g \int_\Gamma \lambda\, v\, ds \le g \int_\Gamma |v|\, ds = j(v),$$

we see that u solves the inequality (11.4.10). □

Another proof of Theorem 11.4.5 is given Exercise 11.4.4.

It is then possible to develop an iterative solution procedure for the inequality problem. Let $\rho > 0$ be a parameter.

INITIALIZATION. Choose $\lambda_0 \in \Lambda$ (e.g. $\lambda_0 = 0$).

ITERATION. For $n = 0, 1, \ldots$, find $u_n \in V$ as the solution of the boundary value problem

$$a(u_n, v) = \ell(v) - g \int_\Gamma \lambda_n v \, ds \quad \forall\, v \in V,$$

and update the Lagrangian multiplier

$$\lambda_{n+1} = \mathcal{P}_\Lambda(\lambda_n + \rho\, g\, u_n).$$

Here $\mathcal{P}_\Lambda$ is a projection operator to Λ defined as

$$\mathcal{P}_\Lambda(\mu) = \sup(-1, \inf(1, \mu)) \quad \forall\, \mu \in L^\infty(\Gamma).$$

It can be shown that there exists a $\rho_0 > 0$ such that if $\rho \in (0, \rho_0)$, then the iterative method converges:

$$u_n \to u \text{ in } V, \quad \lambda_n \to \lambda \text{ in } \Lambda.$$

An interested reader can consult [91, 109] for detailed discussion of the method of Lagrangian multipliers and convergence argument of the iterative method in the context of solving certain other variational inequalities.

Method of numerical integration. We follow [111] to analyze an approach by approximating $j(v_h)$ with $j_h(v_h)$, obtained through numerical integrations. Then the numerical method is

$$u_h \in V_h, \quad (A(u_h), v_h - u_h) + j_h(v_h) - j_h(u_h) \geq \ell(v_h - u_h) \quad \forall\, v_h \in V_h. \tag{11.4.25}$$

For convergence of the numerical method, similar to Theorem 11.4.1, we have the following result; its proof is left as an exercise (see also [91, 93] for the case where A is linear).

Theorem 11.4.6 *Assume $\{V_h\}_h \subset V$ is a family of finite dimensional subspaces such that for a dense subset U of V, one can define mappings $r_h : U \to V_h$ with $\lim_{h\to 0} r_h v = v$ in V, for any $v \in U$. Assume j_h is convex, l.s.c. and uniformly proper in h, and if $v_h \rightharpoonup v$ in V, then $\liminf_{h\to 0} j_h(v_h) \geq j(v)$. Finally, assume $\lim_{h\to 0} j_h(r_h v) = j(v)$ for any $v \in U$. Then for the solution of* (11.4.25)*, we have the convergence*

$$\lim_{h\to 0} \|u - u_h\| = 0.$$

In the above theorem, the functional family $\{j_h\}_h$ is said to be uniformly proper in h, if there exist $\ell_0 \in V^*$ and $c_0 \in \mathbb{R}$ such that

$$j_h(v_h) \geq \ell_0(v_h) + c_0 \quad \forall\, v_h \in V_h,\ \forall\, h.$$

In our application, $j(\cdot)$ is non-negative, as is $j_h(\cdot)$ to be introduced below, so the family $\{j_h\}_h$ is trivially uniformly proper. Notice that Theorem 11.4.6 gives some general assumptions under which one can assert the convergence of the finite element solutions. However, Theorem 11.4.6 does not provide information on the convergence order of the approximations. To derive error estimates we need an inequality of the form (11.4.7).

Theorem 11.4.7 *Assume*

$$j(v_h) \leq j_h(v_h) \quad \forall\, v_h \in V_h. \tag{11.4.26}$$

Let u_h be defined by (11.4.25)*. Then*

$$\|u - u_h\| \leq c \inf_{v_h \in V_h} \left[\|u - v_h\| + |R_h(v_h, u)|^{1/2} \right]. \tag{11.4.27}$$

where

$$R_h(v_h, u) = (A(u), v_h - u) + j_h(v_h) - j(u) - \ell(v_h - u).$$

Proof. Choosing $v = u_h$ in (11.4.11) and adding the resulting inequality to (11.4.25), we obtain

$$\begin{aligned}(A(u), u_h - u) + (A(u_h), v_h - u_h) + j(u_h) - j_h(u_h) + j_h(v_h) - j(u)\\ \geq \ell(v_h - u) \quad \forall\, v_h \in V_h.\end{aligned}$$

Using the assumption (11.4.26) for $v_h = u_h$, we then have

$$(A(u), u_h - u) + (A(u_h), v_h - u_h) + j_h(v_h) - j(u) \geq \ell(v_h - u) \quad \forall\, v_h \in V_h.$$

From this inequality, we obtain

$$(A(u) - A(u_h), u - u_h) \leq (A(u) - A(u_h), u - v_h) + R(v_h, u).$$

We then use the strong monotonicity and Lipschitz continuity of the operator A to derive the bound (11.4.27). □

Let us now comment on the assumption (11.4.26). In some applications, the functional $j(\cdot)$ is of the form $j(v) = I(g\,|v|)$ with I an integration operator, integrating over part or the whole domain or the boundary, $g \geq 0$ is a given non-negative function. One method to construct practically useful approximate functionals j_h is through numerical integrations, $j_h(v_h) = I_h(g\,|v_h|)$. Let $\{\phi_i\}_i$ be the set of functions chosen from a basis of the space

V_h, which defines the functions v_h over the integration region. Assume the basis functions $\{\phi_i\}_i$ are non-negative. Writing

$$v_h = \sum_i v_i \phi_i \quad \text{on the integration region,}$$

we define

$$j_h(v_h) = \sum_i |v_i|\, I(g\,\phi_i). \tag{11.4.28}$$

Obviously the functional $j_h(\cdot)$ constructed in this way enjoys the property (11.4.26). We will see next in the analysis for solving the model problem that certain polynomial invariance property is preserved through a construction of the form (11.4.28). A polynomial invariance property is useful in deriving error estimates.

Let us again consider the model problem (11.4.10). Assume we use linear elements to construct the finite element space V_h. Denote $\{P_i\}$ the set of the nodes of the triangulation which lie on the boundary, numbered consecutively. Let $\{\phi_i\}$ be the canonical basis functions of the space V_h, corresponding to the nodes $\{P_i\}$. Obviously we have the non-negativity property for the basis functions, $\phi_i \geq 0$. Thus according to the formula (11.4.28), we define

$$j_h(v_h) = g \sum_i |\overline{P_i P_{i+1}}|\, \frac{1}{2}\, \left(|v_h(P_i)| + |v_h(P_{i+1})|\right). \tag{11.4.29}$$

Here we use $\overline{P_i P_{i+1}}$ to denote the line segment between P_i and P_{i+1}, and $\overline{P_i P_{i+1}}$ for its length.

Assume $u \in H^2(\Omega)$. Applying Theorem 11.4.7, we have the following bound for the finite element solution error:

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} \leq c \big[& \|u - \Pi_h u\|_{H^1(\Omega)} \\ & + |a(u, \Pi_h u - u) + j_h(\Pi_h u) - j(u) - \ell(\Pi_h u - u)|^{1/2} \big] \end{aligned} \tag{11.4.30}$$

where $\Pi_h u \in V_h$ is the piecewise linear interpolant of the solution u. Let us first estimate the difference $j_h(\Pi_h u) - j(u)$. We have

$$j_h(\Pi_h u) - j(u) = g \sum_i \left\{ \frac{|\overline{P_i P_{i+1}}|}{2} \left[|u(P_i)| + |u(P_{i+1})|\right] - \int_{\overline{P_i P_{i+1}}} |u|\, ds \right\}. \tag{11.4.31}$$

Now if $u|_{\overline{P_iP_{i+1}}}$ keeps the same sign, then

$$\begin{aligned}
&\left| \frac{|\overline{P_iP_{i+1}}|}{2} \left[|u(P_i)| + |u(P_{i+1})| \right] - \int_{\overline{P_iP_{i+1}}} |u| \, ds \right| \\
&\quad = \left| \frac{|\overline{P_iP_{i+1}}|}{2} \left[u(P_i) + u(P_{i+1}) \right] - \int_{\overline{P_iP_{i+1}}} u \, ds \right| \\
&\quad = \left| \int_{\overline{P_iP_{i+1}}} (u - \Pi_h u) \, ds \right| \\
&\quad \le \int_{\overline{P_iP_{i+1}}} |u - \Pi_h u| \, ds.
\end{aligned}$$

Assume $u|_{\overline{P_iP_{i+1}}}$ changes its sign. It is easy to see that

$$\sup_{\overline{P_iP_{i+1}}} |u| \le h \, \|u\|_{W^{1,\infty}(P_iP_{i+1})}$$

if $u|_{\overline{P_iP_{i+1}}} \in W^{1,\infty}(P_iP_{i+1})$, which is valid if $u|_{\Gamma_i} \in H^2(\Gamma_i)$, $i = 1, \ldots, i_0$. Thus,

$$\left| \frac{|\overline{P_iP_{i+1}}|}{2} \left[|u(P_i)| + |u(P_{i+1})| \right] - \int_{\overline{P_iP_{i+1}}} |u| \, ds \right| \le c \, h^2 \|u\|_{W^{1,\infty}(P_iP_{i+1})}.$$

Therefore, if the exact solution u changes its sign only finitely many times on $\partial\Omega$, then from (11.4.31) we find that

$$|j_h(\Pi_h u) - j(u)| \le c \, h^2 \sum_{i=1}^{i_0} \|u\|_{W^{1,\infty}(\Gamma_i)} + c \, \|u - \Pi_h u\|_{L^1(\Gamma)}.$$

Using (11.4.30), we then get

$$\begin{aligned}
\|u - u_h\|_{H^1(\Omega)} \le c \Big\{ & \|u - \Pi_h u\|_{H^1(\Omega)} + \left\| \frac{\partial u}{\partial \nu} \right\|_{L^2(\Gamma)} \|u - \Pi_h u\|_{L^2(\Gamma)} \\
& + h \left[\sum_{i=1}^{i_0} \|u\|_{W^{1,\infty}(\Gamma_i)} \right]^{1/2} + \|u - \Pi_h u\|_{L^1(\Gamma)}^{1/2} \\
& + \| - \Delta u + u - f\|_{L^2(\Omega)} \|u - \Pi_h u\|_{L^2(\Omega)} \Big\}.
\end{aligned}$$

In conclusion, if $u \in H^2(\Omega)$, $u|_{\Gamma_i} \in W^{1,\infty}(\Gamma_i)$ for $i = 1, \ldots, i_0$, and if $u|_\Gamma$ changes its sign only finitely many times, then we have the error estimate

$$\|u - u_h\|_{H^1(\Omega)} \le c(u) \, h,$$

i.e., the approximation of j by j_h does not cause a degradation in the convergence order of the finite element method.

If quadratic elements are used, one can construct basis functions by using nodal shape functions and side modes (see [219]). Then the basis functions are non-negative, and an error analysis similar to the above one can be done.

Exercise 11.4.1 The elasto-plastic torsion problem was introduced in Exercise 11.3.6. Let us consider its one-dimensional analogue and a linear finite element approximation. Let the domain be the unit interval $[0,1]$, and Δ_h be a partition of the interval with the meshsize h. Then the admissible set K_h consists of the continuous piecewise linear functions, vanishing at the ends, and the magnitude of the first derivative being bounded by 1. Show that under suitable solution regularity assumptions, there is an optimal order error estimate $\|u - u_h\|_1 \le c\,h$.

Exercise 11.4.2 Let $\{V_h\}$ be a family of finite dimensional subspaces of V such that $\overline{\cup_h V_h} = V$. In addition to the assumptions made on the data for the variational inequality (11.4.11), we assume further that $j(\cdot)$ is a continuous functional on V. Show that a consequence of (11.4.7) is the convergence of the approximate solutions u_h defined by (11.4.12): $\|u - u_h\| \to 0$ as $h \to 0$.

Exercise 11.4.3 For some choices of the regularization function, there is a constant c_1 such that

$$|j_\varepsilon(v) - j(v)| \le c_1\varepsilon \quad \forall\, v \in V. \tag{11.4.32}$$

Under this additional assumption, derive the error estimate

$$\|u - u_\varepsilon\|_V \le c_2\sqrt{\varepsilon},$$

where $c_2 = \sqrt{2c_1/c_0}$ and c_0 is the V-ellipticity constant of the bilinear form $a(\cdot,\cdot)$. Determine which of the choices from (11.4.15)–(11.4.19) lead to (11.4.32), and identify the corresponding constant c_1 in (11.4.32) when it holds.

Exercise 11.4.4 The equations (11.4.21)–(11.4.22) in Theorem 11.4.5 can be proved via the regularization. Consider (11.4.14)

$$u_\varepsilon \in V, \quad a(u_\varepsilon, v) + \langle j'_\varepsilon(u_\varepsilon), v\rangle = \ell(v) \quad \forall\, v \in V$$

with the regularization function (11.4.15). Denote $\lambda_\varepsilon = \phi'_\varepsilon(u_\varepsilon)$. Then $\lambda_\varepsilon \in \Lambda$ and

$$u_\varepsilon \in V, \quad a(u_\varepsilon, v) + (g\,\lambda_\varepsilon, v)_{L^2(\Gamma)} = \ell(v) \quad \forall\, v \in V. \tag{11.4.33}$$

For some $\lambda \in \Lambda$ and a subsequence $\{\varepsilon\}$,

$$\lambda_\varepsilon \rightharpoonup \lambda \text{ in } L^2(\Omega) \text{ and } \lambda_\varepsilon \rightharpoonup^* \lambda \text{ in } L^\infty(\Omega) \quad \text{as } \varepsilon \to 0.$$

Then take the limit $\varepsilon \to 0$ along the subsequence in (11.4.33).

Carry out the detailed argument.

Exercise 11.4.5 Extend some of the discussions in this section for the numerical analysis of the variational inequalities studied in Exercise 11.3.12.

Exercise 11.4.6 Prove Theorem 11.4.6.

11.5 Some contact problems in elasticity

In this section, we consider some contact problems for elastic bodies, in the form of variational inequalities. It is useful to review the material in Section 8.5 before reading further for this section.

We first introduce some notation and constitutive relations in elasticity, that are needed for studying boundary and initial-boundary value problems arising in contact mechanics.

Let $\boldsymbol{v} : \Omega \to \mathbb{R}^d$ be a vector-valued function from certain Sobolev space. We use the same symbol $\boldsymbol{v}$ to denote the function and its trace on the boundary Γ of Ω. For a vector $\boldsymbol{v}$, we will use its normal component $v_\nu = \boldsymbol{v}\cdot\boldsymbol{\nu}$ and tangential component $\boldsymbol{v}_\tau = \boldsymbol{v} - v_\nu\boldsymbol{\nu}$ at a point on the boundary. Similarly for a tensor-valued function $\boldsymbol{\sigma} : \Omega \to \mathbb{S}^d$, we define its normal component $\sigma_\nu = \boldsymbol{\sigma}\boldsymbol{\nu}\cdot\boldsymbol{\nu}$ and tangential component $\boldsymbol{\sigma}_\tau = \boldsymbol{\sigma}\boldsymbol{\nu} - \sigma_\nu\boldsymbol{\nu}$. Obviously, we have the orthogonality relations $\boldsymbol{v}_\tau\cdot\boldsymbol{\nu} = 0$, $\boldsymbol{\sigma}_\tau\cdot\boldsymbol{\nu} = 0$, and the decomposion formula

$$\boldsymbol{u}\cdot\boldsymbol{v} = (u_\nu\boldsymbol{\nu} + \boldsymbol{u}_\tau)\cdot(v_\nu\boldsymbol{\nu} + \boldsymbol{v}_\tau) = u_\nu v_\nu + \boldsymbol{u}_\tau\cdot\boldsymbol{v}_\tau.$$

For a detailed treatment of traces for vector and tensor fields in contact problems and related spaces see [143] or [115].

We now turn to a description of material constitutive relations. The case of linearized elasticity was discussed in Section 8.5:

$$\boldsymbol{\sigma} = \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u}). \tag{11.5.1}$$

Here and below, for simplicity, we do not specify explicitly the dependence of various functions on $\boldsymbol{x} \in \Omega$. We assume the fourth-order tensor $\mathcal{C}$ is bounded, symmetric and positively definite (pointwise stable) in Ω:

$$\left.\begin{array}{l}
\text{(a) } \mathcal{C} : \Omega\times\mathbb{S}^d \to \mathbb{S}^d.\\
\text{(b) } C_{ijkl} \in L^\infty(\Omega),\ 1\le i,j,k,l\le d.\\
\text{(c) } (\mathcal{C}\boldsymbol{\sigma}):\boldsymbol{\tau} = \boldsymbol{\sigma}:(\mathcal{C}\boldsymbol{\tau}),\ \forall\,\boldsymbol{\sigma},\boldsymbol{\tau}\in\mathbb{S}^d,\ \text{a.e. in } \Omega.\\
\text{(d) There exists } \alpha_0 > 0 \text{ such that}\\
\qquad (\mathcal{C}\boldsymbol{\tau}):\boldsymbol{\tau} \ge \alpha_0|\boldsymbol{\tau}|^2\ \ \forall\,\boldsymbol{\tau}\in\mathbb{S}^d,\ \text{a.e. in } \Omega.
\end{array}\right\} \tag{11.5.2}$$

It can be shown that the condition (11.5.2) (b) is equivalent to

$$C_{ijkl} = C_{klij} = C_{ijlk}, \quad 1\le i,j,k,l\le d.$$

We now describe some constitutive laws for physically nonlinear elastic materials:

$$\boldsymbol{\sigma} = \mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u})) \tag{11.5.3}$$

in which $\mathcal{F}$ is a given nonlinear function. We assume that $\mathcal{F}$ satisfies the following conditions:

$$\left.\begin{array}{l}
\text{(a) } \mathcal{F} : \Omega \times \mathbb{S}^d \to \mathbb{S}^d. \\
\text{(b) There exists } M > 0 \text{ such that} \\
\quad \|\mathcal{F}(\boldsymbol{x}, \boldsymbol{\varepsilon}_1) - \mathcal{F}(\boldsymbol{x}, \boldsymbol{\varepsilon}_2)\| \le M \|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\| \\
\quad\quad \forall\, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_1 \in \mathbb{S}^d, \text{ a.e. } \boldsymbol{x} \in \Omega. \\
\text{(c) There exists } c_0 > 0 \text{ such that} \\
\quad [\mathcal{F}(\boldsymbol{x}, \boldsymbol{\varepsilon}_1) - \mathcal{F}(\boldsymbol{x}, \boldsymbol{\varepsilon}_2)] : (\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2) \ge c_0 \|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|^2 \\
\quad\quad \forall\, \boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_1 \in \mathbb{S}^d, \text{ a.e. } \boldsymbol{x} \in \Omega. \\
\text{(d) For any } \boldsymbol{\varepsilon} \in \mathbb{S}^d, \ \boldsymbol{x} \mapsto \mathcal{F}(\boldsymbol{x}, \boldsymbol{\varepsilon}) \text{ is measurable in } \Omega. \\
\text{(e) The mapping } \boldsymbol{x} \mapsto \mathcal{F}(\boldsymbol{x}, \boldsymbol{0}) \in L^2(\Omega)^{d \times d}.
\end{array}\right\} \tag{11.5.4}$$

Clearly, a family of elasticity operators satisfying the conditions (11.5.4) is provided by the linearly elastic materials (11.5.1), with $\mathcal{F}(\boldsymbol{\varepsilon}) = \mathcal{C}\boldsymbol{\varepsilon}$, if the elasticity tensor satisfies the conditions (11.5.2).

Another example is provided by the nonlinear constitutive law

$$\boldsymbol{\sigma} = \mathcal{C}\boldsymbol{\varepsilon} + \beta\,[\boldsymbol{\varepsilon} - P_K(\boldsymbol{\varepsilon})]\,.$$

Here $\mathcal{C}$ is a fourth-order tensor, $\beta > 0$, K is a closed convex subset of $\mathbb{S}^d$ such that $\boldsymbol{0} \in K$ and $P_K : \mathbb{S}^d \to K$ denotes the projection map. The corresponding elasticity operator is given by

$$\mathcal{F}(\boldsymbol{\varepsilon}) = \mathcal{C}\boldsymbol{\varepsilon} + \beta\,[\boldsymbol{\varepsilon} - P_K(\boldsymbol{\varepsilon})]\,. \tag{11.5.5}$$

Assume the conditions (11.5.2). Due to the nonexpansivity of the projection map, we can verify that the nonlinear operator (11.5.5) satisfies the conditions (11.5.4).

A family of elasticity operators satisfying the conditions (11.5.4) is provided by nonlinear *Hencky materials* (for detail, see e.g., [248]). For a Hencky material, the stress-strain relation is

$$\boldsymbol{\sigma} = K_0 \mathrm{tr}\,\boldsymbol{\varepsilon}(\boldsymbol{u})\,\boldsymbol{I} + \psi(\|\boldsymbol{\varepsilon}^D(\boldsymbol{u})\|^2)\,\boldsymbol{\varepsilon}^D(\boldsymbol{u}),$$

so that the elasticity operator is

$$\mathcal{F}(\boldsymbol{\varepsilon}) = K_0 \mathrm{tr}\,(\boldsymbol{\varepsilon})\,\boldsymbol{I} + \psi(\|\boldsymbol{\varepsilon}^D\|^2)\,\boldsymbol{\varepsilon}^D. \tag{11.5.6}$$

Here, $K_0 > 0$ is a material coefficient, $\boldsymbol{I}$ is the identity tensor of the second order, $\psi : \mathbb{R} \to \mathbb{R}$ is a constitutive function and $\boldsymbol{\varepsilon}^D = \boldsymbol{\varepsilon}^D(\boldsymbol{u})$ denotes the deviatoric part of $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\boldsymbol{u})$:

$$\boldsymbol{\varepsilon}^D = \boldsymbol{\varepsilon} - \frac{1}{d}\,(\mathrm{tr}\,\boldsymbol{\varepsilon})\,\boldsymbol{I}.$$

The function ψ is assumed to be piecewise continuously differentiable, and there exist positive constants c_1, c_2, d_1 and d_2 such that for $\xi \ge 0$,

$$\begin{aligned}
\psi(\xi) &\le d_1, \\
-c_1 \le \psi'(\xi) &\le 0, \\
c_2 \le \psi(\xi) + 2\,\psi'(\xi)\,\xi &\le d_2.
\end{aligned}$$

The conditions (11.5.4) are satisfied for the elasticity operator defined in (11.5.6). For instance, let us verify (11.5.4) (b) and (11.5.4) (c). For $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \in \mathbb{S}^d$ and $t \in [0,1]$, we use the notation

$$\boldsymbol{\varepsilon}^D(t) = \boldsymbol{\varepsilon}_2^D + t\,(\boldsymbol{\varepsilon}_1^D - \boldsymbol{\varepsilon}_2^D).$$

We have

$$\mathcal{F}(\boldsymbol{\varepsilon}_1) - \mathcal{F}(\boldsymbol{\varepsilon}_2) = K_0 \mathrm{tr}(\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2)\,\boldsymbol{I} + \psi(\|\boldsymbol{\varepsilon}_1^D\|^2)\,\boldsymbol{\varepsilon}_1^D - \psi(\|\boldsymbol{\varepsilon}_2^D\|^2)\,\boldsymbol{\varepsilon}_2^D,$$

and

$$\begin{aligned}
\psi(\|\boldsymbol{\varepsilon}_1^D\|^2)\,\boldsymbol{\varepsilon}_1^D - \psi(\|\boldsymbol{\varepsilon}_2^D\|^2)\,\boldsymbol{\varepsilon}_2^D &= \int_0^1 \frac{d}{dt}\big[\psi(\|\boldsymbol{\varepsilon}^D(t)\|^2)\,\boldsymbol{\varepsilon}^D(t)\big]\,dt \\
&= \int_0^1 \big\{2\,\psi'(\|\boldsymbol{\varepsilon}^D(t)\|^2)\,[\boldsymbol{\varepsilon}^D(t) : (\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2)]\,\boldsymbol{\varepsilon}^D(t) \\
&\quad + \psi(\|\boldsymbol{\varepsilon}^D(t)\|^2)\,(\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2)\big\}\,dt.
\end{aligned}$$

Then the condition (11.5.4) (b) is satisfied for some constant M depending on K_0, d_1, d_2 and c_1. Similarly,

$$\begin{aligned}
&[\mathcal{F}(\boldsymbol{\varepsilon}_1) - \mathcal{F}(\boldsymbol{\varepsilon}_2)] : (\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2) \\
&\quad = K_0 |\mathrm{tr}(\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2)|^2 + \int_0^1 \big[2\,\psi'(\|\boldsymbol{\varepsilon}^D(t)\|^2)\,|\boldsymbol{\varepsilon}^D(t) : (\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2)|^2 \\
&\qquad + \psi(\|\boldsymbol{\varepsilon}^D(t)\|^2)\,\|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|^2\big]\,dt \\
&\quad \geq K_0 |\mathrm{tr}(\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2)|^2 \\
&\qquad + \int_0^1 \big[2\,\psi'(\|\boldsymbol{\varepsilon}^D(t)\|^2)\,\|\boldsymbol{\varepsilon}^D(t)\|^2 + \psi(\|\boldsymbol{\varepsilon}^D(t)\|^2)\big]\,\|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|^2\,dt \\
&\quad \geq K_0 |\mathrm{tr}(\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2)|^2 + c_2 \|\boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2\|^2.
\end{aligned}$$

Hence, the condition (11.5.4) (c) is satisfied with m depending on K_0 and c_2.

Remark 11.5.1 In the elastic contact problems to be studied below, we will use the linear constitutive relation (11.5.1) satisfying (11.5.2), or the nonlinear constitutive relation (11.5.3) satisfying (11.5.4). Any discussion involving the nonlinear constitutive relation (11.5.3) can be restated for the linear constitutive relation (11.5.1), and the converse is also true. For this reason, in theoretical analysis of the contact problems, we will not distinguish the two constitutive relations. □

11.5.1 A frictional contact problem

Consider a problem for frictional contact between a linearly elastic body occupying a domain Ω and a rigid foundation. The boundary Γ of Ω is

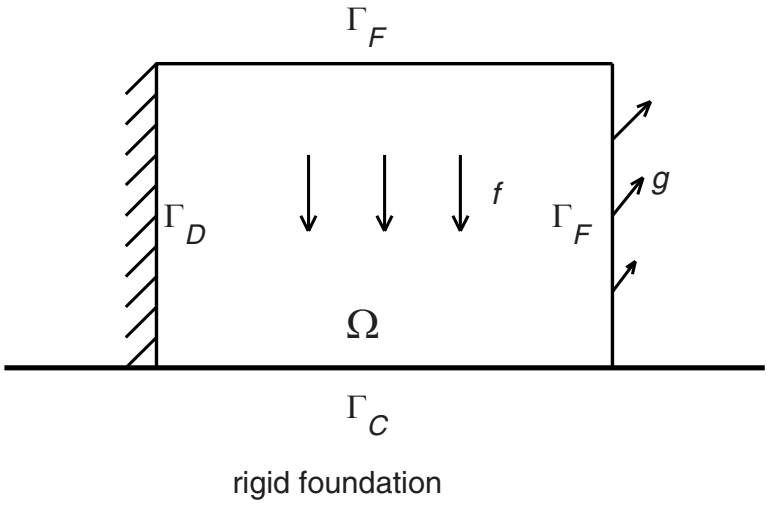

FIGURE 11.2. A body in frictional contact with a rigid foundation

assumed to be Lipschitz continuous and is partitioned into three non-overlapping regions Γ_D, Γ_F and Γ_C where displacement, force (surface traction) and contact boundary conditions are specified, respectively. The body is fixed along Γ_D, and is subject to the action of a body force of the density $\boldsymbol{f}$ and the surface traction of density $\boldsymbol{g}$ on Γ_F. Over Γ_C, the body is in frictional contact with a rigid foundation. The body is assumed to be in equilibrium; see Figure 11.2.

We begin with the specification of the differential equations and boundary conditions. First we have the equilibrium equation

$$-\operatorname{div}\boldsymbol{\sigma} = \boldsymbol{f} \quad \text{in } \Omega, \tag{11.5.7}$$

where $\boldsymbol{\sigma} = (\sigma_{ij})_{d\times d}$ is the stress variable, $\boldsymbol{\sigma} = \boldsymbol{\sigma}^T$. The material is assumed to be linearly elastic with the constitutive relation

$$\boldsymbol{\sigma} = \mathcal{C}\boldsymbol{\varepsilon} \quad \text{in } \Omega, \tag{11.5.8}$$

where $\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}(\boldsymbol{u}) = (\varepsilon_{ij}(\boldsymbol{u}))_{d\times d}$ is the linearized strain tensor

$$\boldsymbol{\varepsilon}(\boldsymbol{u}) = \frac{1}{2}\left[\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^T\right]. \tag{11.5.9}$$

We assume the elasticity tensor $\mathcal{C}$ satisfies the condition (11.5.2).

The boundary conditions on Γ_D and Γ_F are

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_D, \tag{11.5.10}$$

$$\boldsymbol{\sigma}\boldsymbol{\nu} = \boldsymbol{g} \quad \text{on } \Gamma_F. \tag{11.5.11}$$

For boundary conditions on Γ_C, we impose a simplified frictional contact condition ([143, p. 272]):

$$\begin{aligned} &\sigma_\nu = -G, \\ &|\boldsymbol{\sigma}_\tau| \le \mu_F G \quad \text{and} \\ &\quad |\boldsymbol{\sigma}_\tau| < \mu_F G \Longrightarrow \boldsymbol{u}_\tau = \boldsymbol{0}, \\ &\quad |\boldsymbol{\sigma}_\tau| = \mu_F G \Longrightarrow \boldsymbol{u}_\tau = -\lambda\, \boldsymbol{\sigma}_\tau \text{ for some } \lambda \ge 0. \end{aligned} \tag{11.5.12}$$

Here, $G > 0$ and the friction coefficient $\mu_F > 0$ are prescribed functions, $G, \mu_F \in L^\infty(\Gamma_C)$. It is easy to derive from the last two relations of (11.5.12) that

$$\boldsymbol{\sigma}_\tau \cdot \boldsymbol{u}_\tau = -\mu_F G\, |\boldsymbol{u}_\tau| \quad \text{on } \Gamma_C. \tag{11.5.13}$$

The mechanical problem for the frictional contact consists of the relations (11.5.7)–(11.5.12). Let us derive the corresponding weak formulation. We assume the function $\boldsymbol{u}$ is sufficiently smooth so that all the calculations next are valid. We multiply the differential equation (11.5.7) by $\boldsymbol{v} - \boldsymbol{u}$ with an arbitrary $\boldsymbol{v} \in V$,

$$-\int_\Omega \operatorname{div}\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{u})\, dx = \int_\Omega \boldsymbol{f} \cdot (\boldsymbol{v} - \boldsymbol{u})\, dx.$$

Applying the integration by parts formula (8.5.13), using the boundary conditions (11.5.10) and (11.5.11) and the constitutive relation (11.5.8), we obtain

$$\begin{aligned} -\int_\Omega \operatorname{div}\boldsymbol{\sigma} \cdot (\boldsymbol{v} - \boldsymbol{u})\, dx &= -\int_\Gamma \boldsymbol{\sigma}\boldsymbol{\nu} \cdot (\boldsymbol{v} - \boldsymbol{u})\, ds + \int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{v} - \boldsymbol{u})\, dx \\ &= -\int_{\Gamma_F} \boldsymbol{g} \cdot (\boldsymbol{v} - \boldsymbol{u})\, ds - \int_{\Gamma_C} \boldsymbol{\sigma}\boldsymbol{\nu} \cdot (\boldsymbol{v} - \boldsymbol{u})\, ds \\ &\quad + \int_\Omega \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v} - \boldsymbol{u})\, dx. \end{aligned}$$

With the normal and tangential components decompositions of $\boldsymbol{\sigma}\boldsymbol{\nu}$ and $\boldsymbol{v} - \boldsymbol{u}$ on Γ_F, we have

$$\begin{aligned} &-\int_{\Gamma_C} \boldsymbol{\sigma}\boldsymbol{\nu} \cdot (\boldsymbol{v} - \boldsymbol{u})\, ds \\ &\quad = -\int_{\Gamma_C} [\sigma_\nu(v_\nu - u_\nu) + \boldsymbol{\sigma}_\tau \cdot (\boldsymbol{v}_\tau - \boldsymbol{u}_\tau)]\, ds \\ &\quad = \int_{\Gamma_C} G\,(v_\nu - u_\nu)\, ds + \int_{\Gamma_C} (-\boldsymbol{\sigma}_\tau \cdot \boldsymbol{v}_\tau - \mu_F G\, |\boldsymbol{u}_\tau|)\, ds \\ &\quad \le \int_{\Gamma_C} G\,(v_\nu - u_\nu)\, ds + \int_{\Gamma_C} \mu_F G\,(|\boldsymbol{v}_\tau| - |\boldsymbol{u}_\tau|)\, ds, \end{aligned}$$

where the boundary condition (11.5.12) (and its consequence (11.5.13)) is used.

Summarizing, the variational inequality formulation of the problem is to find the displacement field $\boldsymbol{u} \in V$ such that

$$\int_\Omega \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v}-\boldsymbol{u})\,dx + \int_{\Gamma_C} \mu_F G\,|\boldsymbol{v}_\tau|\,ds - \int_{\Gamma_C} \mu_F G\,|\boldsymbol{u}_\tau|\,ds$$
$$\geq \int_\Omega \boldsymbol{f}\cdot(\boldsymbol{v}-\boldsymbol{u})\,dx + \int_{\Gamma_F} \boldsymbol{g}\cdot(\boldsymbol{v}-\boldsymbol{u})\,ds - \int_{\Gamma_C} G\,(v_\nu - u_\nu)\,ds \quad \forall\, \boldsymbol{v} \in V. \tag{11.5.14}$$

Here we assume $\boldsymbol{f} \in [L^2(\Omega)]^d$, $\boldsymbol{g} \in [L^2(\Gamma_F)]^d$. By choosing the function space V for (11.5.14) to be

$$V = \{\boldsymbol{v} \in [H^1(\Omega)]^d \mid \boldsymbol{v} = \boldsymbol{0} \text{ a.e. on } \Gamma_D\},$$

we see that each term in the variational inequality (11.5.14) makes sense.

The corresponding minimization problem is

$$\boldsymbol{u} \in V, \quad E(\boldsymbol{u}) = \inf\{E(\boldsymbol{v}) \mid \boldsymbol{v} \in V\}, \tag{11.5.15}$$

where the energy functional

$$E(\boldsymbol{v}) = \frac{1}{2}\int_\Omega \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{v}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\,dx - \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx$$
$$- \int_{\Gamma_F} \boldsymbol{g}\cdot\boldsymbol{v}\,ds + \int_{\Gamma_C} G\,(v_\nu + \mu_F|\boldsymbol{v}_\tau|)\,ds. \tag{11.5.16}$$

This energy functional is non-differentiable. The non-differentiable term takes the frictional effect into account. The equivalence between the variational inequality (11.5.14) and the minimization problem (11.5.15) is left as Exercise 11.5.1.

It is easy to verify that the conditions stated in Theorem 11.3.7 are satisfied (for the V-ellipticity of the bilinear form, we need to apply Korn's inequality (7.3.12)), and hence the problem has a unique solution.

We now consider numerical approximation of the problem. Assume Ω is a polygon or polyhedron. The boundary is decomposed into three parts $\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_F} \cup \overline{\Gamma_C}$. Write $\overline{\Gamma_C} = \cup_{i=1}^{i_0} \Gamma_{C,i}$ with each $\Gamma_{C,i}$ having a constant outward normal. Let $V_h \subset V$ be the finite element spaces of linear elements corresponding to a regular family of triangulations of $\overline{\Omega}$ that is compatible to the boundary decomposition $\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_F} \cup \cup_{i=1}^{i_0} \Gamma_{C,i}$, i.e., any boundary point common to two sets in this decomposition is a vertex of the finite element triangulations. Then the finite element solution of the variational inequality (11.5.14) is $\boldsymbol{u}_h \in V_h$ such that

$$\int_\Omega \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u}_h) : \boldsymbol{\varepsilon}(\boldsymbol{v}_h - \boldsymbol{u}_h)\,dx + \int_{\Gamma_C} \mu_F G\,|\boldsymbol{v}_{\tau,h}|\,ds - \int_{\Gamma_C} \mu_F G\,|\boldsymbol{u}_{\tau,h}|\,ds$$
$$\geq \int_\Omega \boldsymbol{f}\cdot(\boldsymbol{v}_h - \boldsymbol{u}_h)\,dx + \int_{\Gamma_F} \boldsymbol{g}\cdot(\boldsymbol{v}_h - \boldsymbol{u}_h)\,ds - \int_{\Gamma_C} G\,(v_{\nu,h} - u_{\nu,h})\,ds$$
$$\forall\, \boldsymbol{v}_h \in V_h.$$

The finite element solution exists and is unique.

For error estimation, we need some pointwise equations satisfied by the solution $\boldsymbol{u} \in V$ of (11.5.14). Following the arguments in [115, Section 8.1], we can show that, with $\boldsymbol{\sigma} = \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u})$,

$$-\operatorname{div} \boldsymbol{\sigma} = \boldsymbol{f} \quad \text{a.e. in } \Omega; \tag{11.5.17}$$

moreover, under the assumption

$$\boldsymbol{\sigma\nu} \in L^2(\Gamma)^d, \tag{11.5.18}$$

we also have

$$\boldsymbol{\sigma\nu} = \boldsymbol{g} \quad \text{a.e. on } \Gamma_F, \tag{11.5.19}$$

$$\sigma_\nu = -G \quad \text{a.e. on } \Gamma_C. \tag{11.5.20}$$

The general observation is that under a slightly stronger regularity assumption than $\boldsymbol{u} \in V$ (here it is (11.5.18)), it is possible to show the equalities (here they are (11.5.7), (11.5.11) and $(11.5.12)_1$) of the classical formulation hold a.e. for the solution of the corresponding variational inequality.

Applying Theorem 11.4.2, we have

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V \le c \inf_{\boldsymbol{v}_h \in V_h} \left[\|\boldsymbol{u} - \boldsymbol{v}_h\|_V + |R(\boldsymbol{v}_h, \boldsymbol{u})|^{1/2} \right] \tag{11.5.21}$$

with

$$\begin{aligned} R(\boldsymbol{v}_h, \boldsymbol{u}) = &\int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{v}_h - \boldsymbol{u})\, dx + \int_{\Gamma_C} \mu_F G\, (|\boldsymbol{v}_{\tau,h}| - |\boldsymbol{u}_\tau|)\, ds \\ &- \int_\Omega \boldsymbol{f} \cdot (\boldsymbol{v}_h - \boldsymbol{u})\, dx - \int_{\Gamma_F} \boldsymbol{g} \cdot (\boldsymbol{v}_h - \boldsymbol{u})\, ds \\ &+ \int_{\Gamma_C} G\, (v_{\nu,h} - u_\nu)\, ds. \end{aligned}$$

The inequality (11.5.21) is the basis for convergence analysis and error estimation. It is possible to show the convergence of the numerical solution under the basic solution regularity $\boldsymbol{u} \in V$ ([115, Section 8.2]):

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V \to 0 \quad \text{as } h \to 0.$$

To derive an error bound, with the assumption (11.5.18), we can perform an integration by parts on the first term of $R(\boldsymbol{v}_h, \boldsymbol{u})$ and apply the pointwise relations (11.5.17), (11.5.19) and (11.5.20) to get

$$R(\boldsymbol{v}_h, \boldsymbol{u}) = \int_{\Gamma_C} \left[\mu_F G\, (|\boldsymbol{v}_{\tau,h}| - |\boldsymbol{u}_\tau|) + \boldsymbol{\sigma}_\tau \cdot (\boldsymbol{v}_{\tau,h} - \boldsymbol{u}_\tau) \right] ds.$$

Thus,

$$|R(\boldsymbol{v}_h, \boldsymbol{u})| \le c\, \|\boldsymbol{v}_{\tau,h} - \boldsymbol{u}_\tau\|_{L^2(\Gamma_C)^d}$$

with a constant c depending on $\|\boldsymbol{\sigma}_\tau\|_{L^2(\Gamma_C)^d}$. Then from (11.5.21), we get

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_V \le c \inf_{\boldsymbol{v}_h\in V_h}\left[\|\boldsymbol{u}-\boldsymbol{v}_h\|_V + \|\boldsymbol{v}_{\tau,h}-\boldsymbol{u}_\tau\|^{1/2}_{L^2(\Gamma_C)^d}\right].$$

Under the additional solution regularity conditions

$$\boldsymbol{u}\in H^2(\Omega), \quad u_\nu|_{\Gamma_{C,i}} \in H^2(\Gamma_{C,i}),\ 1\le i\le i_0,$$

we can use the finite element interpolation error estimates to then derive the optimal order error estimate:

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_V = \mathcal{O}(h).$$

11.5.2 A Signorini frictionless contact problem

The physical setting is as follows. An elastic body occupies a domain Ω in $\mathbb{R}^d$ with a Lipschitz boundary Γ that is partitioned into three parts $\overline{\Gamma_D}$, $\overline{\Gamma_F}$ and $\overline{\Gamma_C}$ with Γ_D, Γ_F and Γ_C relatively open and mutually disjoint, such that $\mathrm{meas}\,(\Gamma_D)>0$. The body is clamped on Γ_D and so the displacement field vanishes there. Surface tractions of density $\boldsymbol{g}$ act on Γ_F and volume forces of density $\boldsymbol{f}$ act in Ω. The body is in contact with a rigid foundation on Γ_C. The contact is frictionless and is modeled with the Signorini contact condition in a form with a zero gap function.

Under the previous assumptions, the equilibrium problem of the elastic body in contact with the obstacle can be formulated as follows:

Find a displacement field $\boldsymbol{u}:\Omega\to\mathbb{R}^d$ and a stress field $\boldsymbol{\sigma}:\Omega\to\mathbb{S}^d$ such that

$$\boldsymbol{\sigma} = \mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u})) \qquad \text{in } \Omega, \tag{11.5.22}$$

$$-\mathrm{div}\,\boldsymbol{\sigma} = \boldsymbol{f} \qquad \text{in } \Omega, \tag{11.5.23}$$

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \Gamma_D, \tag{11.5.24}$$

$$\boldsymbol{\sigma\nu} = \boldsymbol{g} \qquad \text{on } \Gamma_F, \tag{11.5.25}$$

$$u_\nu\le 0,\ \sigma_\nu\le 0,\ \sigma_\nu u_\nu = 0,\ \boldsymbol{\sigma}_\tau = \boldsymbol{0} \quad \text{on } \Gamma_C. \tag{11.5.26}$$

To obtain a variational formulation for the mechanical problem, we use the spaces

$$V = \{\,\boldsymbol{v}\in H^1(\Omega)^d \mid \boldsymbol{v}=\boldsymbol{0} \text{ on } \Gamma_D\,\}, \tag{11.5.27}$$

$$Q = \{\,\boldsymbol{\tau}=(\tau_{ij})\in L^2(\Omega)^{d\times d} \mid \tau_{ij}=\tau_{ji},\ 1\le i,j\le d\,\}. \tag{11.5.28}$$

These are real Hilbert space with their canonical inner products. Over the space V, we use the inner product

$$(\boldsymbol{u},\boldsymbol{v})_V = (\boldsymbol{\varepsilon}(\boldsymbol{u}),\boldsymbol{\varepsilon}(\boldsymbol{v}))_Q \quad \forall\,\boldsymbol{u},\boldsymbol{v}\in V.$$

We use

$$K = \{\boldsymbol{v} \in V \mid v_\nu \leq 0 \text{ a.e. on } \Gamma_C\}$$

as the set of admissible displacement fields, which is a non-empty closed convex set of V.

We assume that the elasticity operator $\mathcal{F}$ satisfies the condition (11.5.4). We also assume that the force and traction densities satisfy

$$\boldsymbol{f} \in L^2(\Omega)^d, \quad \boldsymbol{g} \in L^2(\Gamma_F)^d \tag{11.5.29}$$

and we define $\ell \in V'$ by

$$\ell(\boldsymbol{v}) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}\, dx + \int_{\Gamma_F} \boldsymbol{g} \cdot \boldsymbol{v}\, ds \quad \forall\, \boldsymbol{v} \in V. \tag{11.5.30}$$

We follow the standard procedure to derive the weak formulation of the mechanical problem (11.5.22)–(11.5.26). Assume the mechanical problem has a solution $(\boldsymbol{u}, \boldsymbol{\sigma})$, sufficiently smooth so that all the calculations below are meaningful. Let $\boldsymbol{v} \in K$ be arbitrary. We multiply the equation (11.5.23) by $(\boldsymbol{v} - \boldsymbol{u})$, integrate over Ω, and integrate by parts to obtain

$$(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{v}) - \boldsymbol{\varepsilon}(\boldsymbol{u}))_Q = \int_\Omega \boldsymbol{f} \cdot (\boldsymbol{v} - \boldsymbol{u})\, dx + \int_\Gamma \boldsymbol{\sigma\nu} \cdot (\boldsymbol{v} - \boldsymbol{u})\, ds.$$

Using the boundary conditions (11.5.24) and (11.5.25), we deduce the equality

$$(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{v}) - \boldsymbol{\varepsilon}(\boldsymbol{u}))_Q = \ell(\boldsymbol{v} - \boldsymbol{u}) + \int_{\Gamma_C} \boldsymbol{\sigma\nu} \cdot (\boldsymbol{v} - \boldsymbol{u})\, ds. \tag{11.5.31}$$

Over Γ_C,

$$\boldsymbol{\sigma\nu} \cdot (\boldsymbol{v} - \boldsymbol{u}) = \sigma_\nu(v_\nu - u_\nu) + \boldsymbol{\sigma}_\tau \cdot (\boldsymbol{v}_\tau - \boldsymbol{u}_\tau) = \sigma_\nu v_\nu,$$

where the boundary conditions (11.5.26) are used. Now over Γ_C, we also have $\sigma_\nu \leq 0$ and $v_\nu \leq 0$. Then

$$\boldsymbol{\sigma\nu} \cdot (\boldsymbol{v} - \boldsymbol{u}) \geq 0 \quad \text{on } \Gamma_C.$$

Therefore, we derive from (11.5.31) that

$$(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{v}) - \boldsymbol{\varepsilon}(\boldsymbol{u}))_Q \geq \ell(\boldsymbol{v} - \boldsymbol{u}). \tag{11.5.32}$$

By the boundary conditions (11.5.24) and (11.5.26),

$$\boldsymbol{u} \in K. \tag{11.5.33}$$

Summarizing, from (11.5.22), (11.5.32) and (11.5.33) we arrive at the following variational formulation of the contact problem (11.5.22)–(11.5.26): Find a displacement field $\boldsymbol{u}$ such that

$$\boldsymbol{u} \in K, \quad (\mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u})), \boldsymbol{\varepsilon}(\boldsymbol{v}) - \boldsymbol{\varepsilon}(\boldsymbol{u}))_Q \geq \ell(\boldsymbol{v} - \boldsymbol{u}) \quad \forall\, \boldsymbol{v} \in K. \tag{11.5.34}$$

Applying Theorem 11.3.1, we find that assuming (11.5.4) and (11.5.29), there exists a unique solution $\boldsymbol{u} \in K$ to the variational inequality (11.5.34). Moreover, the mapping $(\boldsymbol{f}, \boldsymbol{g}) \mapsto \boldsymbol{u}$ is Lipschitz continuous from $L^2(\Omega)^d \times L^2(\Gamma_F)^d$ to V.

As in Subsection 11.5.1, we have the equality

$$-\operatorname{div} \boldsymbol{\sigma} = \boldsymbol{f} \quad \text{a.e. in } \Omega, \tag{11.5.35}$$

and under the regularity assumption

$$\boldsymbol{\sigma\nu} \in L^2(\Gamma)^d, \tag{11.5.36}$$

it is possible to establish the following pointwise relations for the solution:

$$\boldsymbol{\sigma\nu} = \boldsymbol{g} \quad \text{a.e. on } \Gamma_F, \tag{11.5.37}$$

$$\boldsymbol{\sigma}_\tau = \boldsymbol{0} \quad \text{a.e. on } \Gamma_C. \tag{11.5.38}$$

We now consider numerical approximation of the problem. Assume Ω is a polygon or polyhedron. Let $V_h \subset V$ be the finite element spaces of linear elements corresponding to a regular family of triangulations of $\overline{\Omega}$ that is compatible to the boundary decomposition $\Gamma = \overline{\Gamma_D} \cup \overline{\Gamma_F} \cup \cup_{i=1}^{i_0} \Gamma_{C,i}$. Let $K_h = V_h \cap K$ be the finite element subset of K. An element from K is a continuous piecewise linear function with vanishing values at the finite element nodes lying on $\overline{\Gamma_D}$ such that its normal component at any node on $\overline{\Gamma_C}$ is non-positive. The discrete approximation of the variational inequality (11.5.34) is to find a displacement field $\boldsymbol{u}_h$ such that

$$\boldsymbol{u}_h \in K_h, \quad (\mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u}_h)), \boldsymbol{\varepsilon}(\boldsymbol{v}_h) - \boldsymbol{\varepsilon}(\boldsymbol{u}_h))_Q \geq \ell(\boldsymbol{v}_h - \boldsymbol{u}_h) \quad \forall\, \boldsymbol{v}_h \in K_h. \tag{11.5.39}$$

With the assumptions made on the data, the discrete variational inequality (11.5.39) has a unique solution. Here we are interested in estimating the error $\boldsymbol{u} - \boldsymbol{u}_h$. Notice that $K_h \subset K$. Applying Theorem 11.4.2, we have

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V \leq c \inf_{\boldsymbol{v}_h \in K_h} \left[\|\boldsymbol{u} - \boldsymbol{v}_h\|_V + |R(\boldsymbol{v}_h, \boldsymbol{u})|^{1/2} \right]. \tag{11.5.40}$$

where

$$R(\boldsymbol{v}_h, \boldsymbol{u}) = \big(\mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u})), \boldsymbol{\varepsilon}(\boldsymbol{v}_h) - \boldsymbol{\varepsilon}(\boldsymbol{u})\big)_Q - \ell(\boldsymbol{v}_h - \boldsymbol{u}). \tag{11.5.41}$$

The inequality (11.5.40) is the basis for convergence analysis and error estimation. It is possible to show the convergence of the numerical solution under the basic solution regularity $\boldsymbol{u} \in K$ ([115, Section 8.2]):

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V \to 0 \quad \text{as } h \to 0.$$

We now derive an error estimate. We need to bound the residual term

$$R(\boldsymbol{v}_h, \boldsymbol{u}) = \big(\boldsymbol{\sigma}, \boldsymbol{\varepsilon}(\boldsymbol{v}_h) - \boldsymbol{\varepsilon}(\boldsymbol{u})\big)_Q - \ell(\boldsymbol{v}_h - \boldsymbol{u})_V.$$

Integrate by parts on the first term and use the pointwise relation (11.5.35),

$$R(\boldsymbol{v}_h, \boldsymbol{u}) = \langle \boldsymbol{\sigma\nu}, \boldsymbol{v}_h - \boldsymbol{u} \rangle_\Gamma - \int_{\Gamma_F} \boldsymbol{g} \cdot (\boldsymbol{v}_h - \boldsymbol{u})\, ds.$$

Under the regularity assumption (11.5.36), we can use the pointwise relations (11.5.37) and (11.5.38) to get

$$R(\boldsymbol{v}_h, \boldsymbol{u}) = \int_{\Gamma_C} \sigma_\nu (v_{\nu,h} - u_\nu)\, ds.$$

Then we obtain

$$|R(\boldsymbol{v}_h, \boldsymbol{u})| \le \|\sigma_\nu\|_{L^2(\Gamma_C)} \|v_{\nu,h} - u_\nu\|_{L^2(\Gamma_C)},$$

and (11.5.40) is reduced to

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V \le c \inf_{\boldsymbol{v}_h \in K_h} \left[\|\boldsymbol{u} - \boldsymbol{v}_h\|_V + \|v_{\nu,h} - u_\nu\|^{1/2}_{L^2(\Gamma_C)} \right]. \tag{11.5.42}$$

Here we allow the constant c to depend on $\|\sigma_\nu\|_{L^2(\Gamma_C)}$.

Then under the additional regularity assumptions

$$\boldsymbol{u} \in H^2(\Omega), \quad u_\nu|_{\Gamma_{C,i}} \in H^2(\Gamma_{C,i}),\ 1 \le i \le i_0,$$

we can use the finite element interpolation error estimates to obtain the optimal order error estimate from (11.5.42):

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_V = \mathcal{O}(h).$$

Exercise 11.5.1 Prove that the variational inequality (11.5.14) and the minimization problem (11.5.15) are mutually equivalent.

Exercise 11.5.2 As another frictional contact problem, we keep the equations and relations (11.5.7)–(11.5.11), but replace (11.5.12) by the following relations on Γ_C:

$$u_\nu = 0, \tag{11.5.43}$$
$$|\boldsymbol{\sigma}_\tau| \le g, \tag{11.5.44}$$
$$|\boldsymbol{\sigma}_\tau| < g \ \Rightarrow\ \boldsymbol{u}_\tau = \boldsymbol{0}, \tag{11.5.45}$$
$$|\boldsymbol{\sigma}_\tau| = g \ \Rightarrow\ \boldsymbol{u}_\tau = -\kappa\, \boldsymbol{\sigma}_\tau \text{ for some } \kappa \ge 0. \tag{11.5.46}$$

Here (11.5.43) is the assumption that the body is in bilateral contact with the rigid foundation; the contact is frictional and is modeled by Tresca's law (11.5.44)–(11.5.46), the friction bound $g > 0$.

Let

$$V = \left\{ \boldsymbol{v} \in \left[H^1(\Omega)\right]^d \mid \boldsymbol{v}|_{\Gamma_D} = \boldsymbol{0},\ v_\nu|_{\Gamma_C} = 0 \right\}$$

with its inner product and norm defined by

$$(\boldsymbol{u},\boldsymbol{v})_V = \int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\,dx, \quad \|\boldsymbol{v}\|_V = (\boldsymbol{v},\boldsymbol{v})_V^{1/2}.$$

Derive the following weak formulation of the contact problem:

$$\boldsymbol{u} \in V, \quad a(\boldsymbol{u},\boldsymbol{v}-\boldsymbol{u}) + j(\boldsymbol{v}) - j(\boldsymbol{u}) \geq \ell(\boldsymbol{v}-\boldsymbol{u}) \quad \forall\, \boldsymbol{v} \in V, \tag{11.5.47}$$

where

$$\begin{aligned} a(\boldsymbol{u},\boldsymbol{v}) &= \int_\Omega \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\,dx, \\ \ell(\boldsymbol{v}) &= \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}\,dx + \int_{\Gamma_N} \boldsymbol{g}\cdot\boldsymbol{v}\,ds, \\ j(\boldsymbol{v}) &= \int_{\Gamma_C} g\,|\boldsymbol{v}_\tau|\,ds. \end{aligned}$$

Show that the variational inequality (11.5.47) has a unique solution $\boldsymbol{u} \in V$, which is the minimizer of the energy functional

$$E(\boldsymbol{v}) = \frac{1}{2}a(\boldsymbol{v},\boldsymbol{v}) + j(\boldsymbol{v}) - \ell(\boldsymbol{v})$$

over the space V. Define a numerical method for solving the variational inequality and derive error estimates.

Exercise 11.5.3 In this exercise, we study a frictionless contact problem with deformable support. This problem differs from the one studied in this section in that the support is not rigid but elastic. We use a version of the normal compliance condition, in a form with a zero gap function, to describe the reaction of the foundation when there is penetration, $u_\nu > 0$. We suppose that the normal stress on the contact surface is proportional to a power α of the penetration depth of the elastic body into the obstacle.

The classical formulation of the problem is to find a displacement field $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ and a stress field $\boldsymbol{\sigma} : \Omega \to \mathbb{S}^d$ such that (11.5.22)–(11.5.25) are satisfied together with

$$-\sigma_\nu = \mu\,(u_\nu)_+^\alpha, \ \boldsymbol{\sigma}_\tau = \boldsymbol{0} \text{ on } \ \Gamma_C. \tag{11.5.48}$$

Here $\mu > 0$ is a penalization parameter, which may be interpreted as a *deformability coefficient* of the foundation, $r_+ = \max\{0, r\}$ and $0 < \alpha \leq 1$. Notice that formally (11.5.48) becomes (11.5.26) when $\mu \to 0+$.

(a) Use the spaces V and Q defined in (11.5.27) and (11.5.28), respectively, and assume the conditions (11.5.4) and (11.5.29). Define $\ell \in V'$ by (11.5.30) and denote by $j : V \to \mathbb{R}$ the functional

$$j(\boldsymbol{v}) = \frac{1}{\mu\,(\alpha+1)} \int_{\Gamma_C} (v_\nu)_+^{\alpha+1} ds \quad \forall\, \boldsymbol{v} \in V.$$

Derive the following variational formulation of the contact problem (11.5.22)–(11.5.25) and (11.5.48): Find a displacement field $\boldsymbol{u} \in V$ such that

$$(\mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u})), \boldsymbol{\varepsilon}(\boldsymbol{v}) - \boldsymbol{\varepsilon}(\boldsymbol{u}))_Q + j(\boldsymbol{v}) - j(\boldsymbol{u}) \geq \ell(\boldsymbol{v}-\boldsymbol{u}) \quad \forall\, \boldsymbol{v} \in V. \tag{11.5.49}$$

Hint: Show that if $(\boldsymbol{u},\boldsymbol{\sigma})$ is a regular solution of the contact problem, then for any $\boldsymbol{v}\in V$,

$$\boldsymbol{\sigma\nu}\cdot(\boldsymbol{v}-\boldsymbol{u})\geq\frac{1}{\mu\,(\alpha+1)}(u_\nu)_+^{\alpha+1}-\frac{1}{\mu\,(\alpha+1)}(v_\nu)_+^{\alpha+1}\quad\text{a.e. on }\Gamma_C.$$

(b) Show that under the assumptions (11.5.4) and (11.5.29), the variational inequality (11.5.49) has a unique solution $\boldsymbol{u}$, and the mapping $(\boldsymbol{f},\boldsymbol{g})\mapsto\boldsymbol{u}$ is Lipschitz continuous from $L^2(\Omega)^d\times L^2(\Gamma_F)$ to V.
Remark. Denote the solution of (11.5.49) by $(\boldsymbol{u}_\mu,\boldsymbol{\sigma}_\mu)$, and the solution of (11.5.34) by $(\boldsymbol{u},\boldsymbol{\sigma})$. It is proved in [70] that as $\mu\to 0$,

$$\boldsymbol{u}_\mu\to\boldsymbol{u}\text{ in }V,\quad\boldsymbol{\sigma}_\mu\to\boldsymbol{\sigma}\text{ in }Q,\quad\mathrm{div}\boldsymbol{\sigma}_\mu\to\mathrm{div}\boldsymbol{\sigma}\text{ in }L^2(\Omega)^d.$$

(c) Let $V_h\subset V$ be the finite element space of linear elements used in this section and introduce the discrete approximation problem

$$\begin{aligned}\boldsymbol{u}_h\in V_h,\quad&(\mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u}_h)),\boldsymbol{\varepsilon}(\boldsymbol{v}_h)-\boldsymbol{\varepsilon}(\boldsymbol{u}_h))_Q+j(\boldsymbol{v}_h)-j(\boldsymbol{u}_h)\\&\geq\ell(\boldsymbol{v}_h-\boldsymbol{u}_h)_V\quad\forall\,\boldsymbol{v}_h\in V_h.\end{aligned}$$

Show that this problem has a unique solution, and for the error $\boldsymbol{u}-\boldsymbol{u}_h$, we have

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_V\leq c\inf_{\boldsymbol{v}_h\in V_h}\left[\|\boldsymbol{u}-\boldsymbol{v}_h\|_V+|R(\boldsymbol{v}_h,\boldsymbol{u})|^{1/2}\right],\tag{11.5.50}$$

where

$$R(\boldsymbol{v}_h,\boldsymbol{u})=\big(\mathcal{F}(\boldsymbol{\varepsilon}(\boldsymbol{u})),\boldsymbol{\varepsilon}(\boldsymbol{v}_h)-\boldsymbol{\varepsilon}(\boldsymbol{u})\big)_Q+j(\boldsymbol{v}_h)-j(\boldsymbol{u})-\ell(\boldsymbol{v}_h-\boldsymbol{u}).\tag{11.5.51}$$

Under the regularity assumption (11.5.36), derive the error bound

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_V\leq c\inf_{\boldsymbol{v}_h\in V_h}\left[\|\boldsymbol{u}-\boldsymbol{v}_h\|_V+\|v_{\nu,h}-u_\nu\|_{L^2(\Gamma_C)}^{1/2}+\|v_{\nu,h}-u_\nu\|_{L^2(\Gamma_C)}^{(1+\alpha)/2}\right].$$

Here we allow the constant c to depend on the solution $(\boldsymbol{u},\boldsymbol{\sigma})$. Then with the additional regularity assumptions

$$\boldsymbol{u}\in H^2(\Omega),\quad u_\nu|_{\Gamma_{C,i}}\in H^2(\Gamma_{C,i}),\ 1\leq i\leq i_0,$$

prove the optimal order error estimate

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_V=\mathcal{O}(h).$$

Remark. It follows from the inequality (11.5.50) that under the basic solution regularity $\boldsymbol{u}\in V$,

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|_V\to 0\quad\text{as }h\to 0.$$

Suggestion for Further Reading.

Interest on variational inequalities originates in mechanical problems. An early reference is Fichera [81]. The first rigorous comprehensive mathematical treatment seems to be Lions and Stampacchia [159]. Duvaut and Lions [74] formulated and studied many problems in mechanics and physics in the framework of variational inequalities. More recent references

include Friedman [85] (mathematical analysis of various variational inequalities in mechanics), Glowinski [91] and Glowinski, Lions and Trémolières [93] (numerical analysis and solution algorithms), Han and Reddy [113] (mathematical and numerical analysis of variational inequalities arising in hardening plasticity), Han and Sofonea [115] (variational and numerical analysis of variational inequalities in contact mechanics for viscoelastic and viscoplastic materials), Haslinger, Hlaváček and Nečas [119] and Hlaváček, Haslinger, Nečas and Lovíšek [126] (numerical solution of variational inequalities in mechanics), Kikuchi and Oden [143] (numerical analysis of various contact problems in elasticity), Kinderlehrer and Stampacchia [144] (a mathematical introduction to the theory of variational inequalities), Panagiotopoulos [185] (theory and numerical approximations of variational inequalities in mechanics), and Sofonea and Matei [208] (mathematical theory of antiplane frictional contact problems).

In the numerical solution of variational inequalities of higher-order, nonconforming finite element methods offer a great advantage. Error analysis for some non-conforming finite element methods in solving an EVI of the first kind arising in unilateral problem, and an EVI of the second kind arising in a plate contact problem can be found in [228] and [116], respectively.

12

Numerical Solution of Fredholm Integral Equations of the Second Kind

Linear integral equations of the second kind,

$$\lambda u(x) - \int_D k(x,y)\, u(y)\, dy = f(x), \quad x \in D \tag{12.0.1}$$

were introduced in Chapter 2, and we note that they occur in a wide variety of physical applications. An important class of such equations are the *boundary integral equations*, about which more is said in Chapter 13. In the integral of (12.0.1), D is a *closed*, and often bounded, integration region. The integral operator is often a compact operator on $C(D)$ or $L^2(D)$, although not always. For the case that the integral operator is compact, a general solvability theory is given in Subsection 2.8.4 of Chapter 2. A more general introduction to the theory of such equations is given in Kress [149].

In this chapter, we look at the two most important classes of numerical methods for these equations: *projection methods* and *Nyström methods*. In Section 12.1, we introduce collocation and Galerkin methods, beginning with explicit definitions and followed by an abstract framework for the analysis of all projection methods. Illustrative examples are given in Section 12.2, and the *iterated projection method* is defined, analyzed, and illustrated in Section 12.3. The Nyström method is introduced and discussed in Section 12.4, and it is extended to the use of product integration in Section 12.5. In Section 12.6, we introduce and analyze some iteration methods. We conclude the chapter in Section 12.7 by introducing and analyzing projection methods for solving some fixed point problems for nonlinear operators.

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_12,
© Springer Science + Business Media, LLC 2009

In this chapter, we use notation that is popular in the literature on the numerical solution of integral equations. For example, the spatial variable is denoted by x, not $\boldsymbol{x}$, in the multi-dimensional case.

12.1 Projection methods: General theory

With all projection methods, we consider solving (12.0.1) within the framework of some complete function space V, usually $C(D)$ or $L^2(D)$. We choose a sequence of finite dimensional approximating subspaces $V_n \subset V$, $n \geq 1$, with V_n having dimension κ_n. Let V_n have a basis $\{\phi_1, \ldots, \phi_\kappa\}$, with $\kappa \equiv \kappa_n$ for notational simplicity (which is done at various points throughout the chapter). We seek a function $u_n \in V_n$, which can be written as

$$u_n(x) = \sum_{j=1}^{\kappa_n} c_j \phi_j(x), \quad x \in D. \tag{12.1.1}$$

This is substituted into (12.0.1), and the coefficients $\{c_1, \ldots, c_\kappa\}$ are determined by forcing the equation to be almost exact in some sense. For later use, introduce

$$\begin{aligned} r_n(x) &= \lambda u_n(x) - \int_D k(x,y) u_n(y)\, dy - f(x) \\ &= \sum_{j=1}^{\kappa} c_j \left[\lambda \phi_j(x) - \int_D k(x,y)\phi_j(y)\, dy \right] - f(x), \end{aligned} \tag{12.1.2}$$

for $x \in D$. This quantity is called the *residual* in the approximation of the equation when using $u \approx u_n$. As usual, we write (12.0.1) in operator notation as

$$(\lambda - K)\, u = f. \tag{12.1.3}$$

Then the residual can be written as

$$r_n = (\lambda - K)\, u_n - f.$$

The coefficients $\{c_1, \ldots, c_\kappa\}$ are chosen by forcing $r_n(x)$ to be approximately zero in some sense. The hope, and expectation, is that the resulting function $u_n(x)$ will be a good approximation of the true solution $u(x)$.

12.1.1 Collocation methods

Pick distinct node points $x_1, \ldots, x_\kappa \in D$, and require

$$r_n(x_i) = 0, \quad i = 1, \ldots, \kappa_n. \tag{12.1.4}$$

This leads to determining $\{c_1, \dots, c_\kappa\}$ as the solution of the linear system

$$\sum_{j=1}^{\kappa} c_j \left[\lambda \phi_j(x_i) - \int_D k(x_i, y) \phi_j(y)\, dy \right] = f(x_i), \quad i = 1, \dots, \kappa. \tag{12.1.5}$$

An immediate question is whether this system has a solution and whether it is unique. If so, does u_n converge to u? Note also that the linear system contains integrals which must usually be evaluated numerically, a point we return to later. We should have written the node points as $\{x_{1,n}, \dots, x_{\kappa,n}\}$; but for notational simplicity, the explicit dependence on n has been suppressed, to be understood only implicitly.

The function space framework for collocation methods is often $C(D)$, which is what we use here. It is possible to use extensions of $C(D)$. For example, we can use $L^\infty(D)$, making use of the ideas of Example 2.5.3 from Section 2.5 to extend the idea of point evaluation of a continuous function to elements of $L^\infty(D)$. Such extensions of $C(D)$ are needed when the approximating functions u_n are not required to be continuous.

As a part of writing (12.1.5) in a more abstract form, we introduce a projection operator P_n which maps $V = C(D)$ onto V_n. Given $u \in C(D)$, define $P_n u$ to be that element of V_n which interpolates u at the nodes $\{x_1, \dots, x_\kappa\}$. This means writing

$$P_n u(x) = \sum_{j=1}^{\kappa_n} \alpha_j \phi_j(x)$$

with the coefficients $\{\alpha_j\}$ determined by solving the linear system

$$\sum_{j=1}^{\kappa_n} \alpha_j \phi_j(x_i) = u(x_i), \quad i = 1, \dots, \kappa_n.$$

This linear system has a unique solution if

$$\det\left(\phi_j(x_i)\right) \neq 0. \tag{12.1.6}$$

Henceforth in this chapter, we assume this is true whenever the collocation method is being discussed. By a simple argument, this condition also implies that the functions $\{\phi_1, \dots, \phi_\kappa\}$ are a linearly independent set over D. In the case of polynomial interpolation for functions of one variable and monomials $\{1, x, \dots, x^n\}$ as the basis functions, the determinant in (12.1.6) is referred to as the *Vandermonde determinant*; see (3.2.2) for its value.

To see more clearly that P_n is linear, and to give a more explicit formula, we introduce a new set of basis functions. For each i, $1 \le i \le \kappa_n$, let $\ell_i \in V_n$ be that element which satisfies the interpolation conditions

$$\ell_i(x_j) = \delta_{ij}, \quad j = 1, \dots, \kappa_n. \tag{12.1.7}$$

By (12.1.6), there is a unique such ℓ_i; and the set $\{\ell_1, \ldots, \ell_\kappa\}$ is a new basis for V_n. With polynomial interpolation, such functions ℓ_i are called *Lagrange basis functions*; and we use this name with all types of approximating subspaces V_n. With this new basis, we can write

$$P_n u(x) = \sum_{j=1}^{\kappa_n} u(x_j)\ell_j(x), \quad x \in D. \tag{12.1.8}$$

Recall (3.2.1)–(3.2.3) in Chapter 3. Clearly, P_n is linear and finite rank. In addition, as an operator on $C(D)$ to $C(D)$,

$$\|P_n\| = \max_{x\in D} \sum_{j=1}^{\kappa_n} |\ell_j(x)|. \tag{12.1.9}$$

Example 12.1.1 Let $V_n = \text{span}\{1, x, \ldots, x^n\}$. Recall the Lagrange interpolatory projection operator of Example 3.6.5:

$$P_n g(x) \equiv \sum_{i=0}^{n} g(x_i)\ell_i(x) \tag{12.1.10}$$

with the Lagrange basis functions

$$\ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j}\right), \quad i = 0, 1, \ldots, n.$$

This is *Lagrange's form of the interpolation polynomial*. In Subsection 3.7.3 of Chapter 3, we denoted this projection operator by $\mathcal{I}_n$. □

Returning to (12.1.8), we note that

$$P_n g = 0 \quad \text{if and only if} \quad g(x_j) = 0,\ j = 1, \ldots, \kappa_n. \tag{12.1.11}$$

The condition (12.1.5) can now be rewritten as

$$P_n r_n = 0$$

or equivalently,

$$P_n(\lambda - K)u_n = P_n f, \quad u_n \in V_n. \tag{12.1.12}$$

We return to this below.

12.1.2 Galerkin methods

Let $V = L^2(D)$ or some other Hilbert function space, and let $(\cdot, \cdot)$ denote the inner product for V. Require r_n to satisfy

$$(r_n, \phi_i) = 0, \quad i = 1, \ldots, \kappa_n. \tag{12.1.13}$$

The left side is the Fourier coefficient of r_n associated with ϕ_i. If $\{\phi_1, \ldots, \phi_\kappa\}$ consists of the leading members of an orthonormal family $\Phi \equiv \{\phi_i\}_{i\geq 1}$ which spans V, then (12.1.13) requires the leading terms to be zero in the Fourier expansion of r_n with respect to Φ.

To find u_n, apply (12.1.13) to (12.0.1) written as $(\lambda - K)\, u = f$. This yields the linear system

$$\sum_{j=1}^{\kappa_n} c_j \left[\lambda(\phi_j, \phi_i) - (K\phi_j, \phi_i)\right] = (f, \phi_i), \quad i = 1, \ldots, \kappa_n. \tag{12.1.14}$$

This is Galerkin's method for obtaining an approximate solution to (12.0.1) or (12.1.3). Does the system have a solution? If so, is it unique? Does the resulting sequence of approximate solutions u_n converge to u in V? Does the sequence converge in $C(D)$, i.e., does u_n converge uniformly to u? Note also that the above formulation contains double integrals $(K\phi_j, \phi_i)$. These must often be computed numerically; and later, we return to a consideration of this.

As a part of writing (12.1.14) in a more abstract form, we recall the orthogonal projection operator P_n of Proposition 3.6.9 of Section 3.6 in Chapter 3, which maps V onto V_n. Recall that

$$P_n g = 0 \quad \text{if and only if} \quad (g, \phi_i) = 0,\ i = 1, \ldots, \kappa_n. \tag{12.1.15}$$

With P_n, we can rewrite (12.1.13) as

$$P_n r_n = 0,$$

or equivalently,

$$P_n(\lambda - K)u_n = P_n f, \qquad u_n \in V_n. \tag{12.1.16}$$

Note the similarity to (12.1.12).

There is a variant on Galerkin's method, known as the *Petrov-Galerkin method* (see Section 9.2). With it, we still choose $u_n \in V_n$; but now we require

$$(r_n, w) = 0 \quad \forall\, w \in W_n$$

with W_n another finite dimensional subspace, also of dimension κ_n. This method is not considered further in this chapter; but it is an important method when we look at the numerical solution of boundary integral equations. Another theoretical approach to Galerkin's method is to set it within a "variational framework", which is done in Chapter 10 and leads to finite element methods.

12.1.3 A general theoretical framework

Let V be a Banach space, and let $\{V_n \mid n \geq 1\}$ be a sequence of finite dimensional subspaces. Denote the dimension of V_n by κ_n, and we assume

$\kappa_n \to \infty$ as $n \to \infty$. Let $P_n : V \to V_n$ be a bounded projection operator. This means that P_n is a bounded linear operator with the property

$$P_n v = v \quad \forall\, v \in V_n.$$

Note that this property implies $P_n^2 = P_n$, and thus

$$\|P_n\| = \|P_n^2\| \le \|P_n\|^2.$$

Therefore,

$$\|P_n\| \ge 1. \tag{12.1.17}$$

Recall the earlier discussion of projection operators from Section 3.6 in Chapter 3. We already have examples of P_n in the interpolatory projection operator and the orthogonal projection operator introduced above in defining the collocation and Galerkin methods, respectively.

Motivated by (12.1.12) and (12.1.16), we now approximate the equation (12.1.3), $(\lambda - K)\, u = f$, by attempting to solve the problem

$$P_n(\lambda - K)\, u_n = P_n f, \quad u_n \in V_n. \tag{12.1.18}$$

This is the form in which the method is implemented, as it leads directly to equivalent finite linear systems such as (12.1.5) and (12.1.14). For the error analysis, however, we write (12.1.18) in an equivalent but more convenient form.

If u_n is a solution of (12.1.18), then by using $P_n u_n = u_n$, the equation can be written as

$$(\lambda - P_n K)\, u_n = P_n f, \quad u_n \in V. \tag{12.1.19}$$

To see that a solution of this is also a solution of (12.1.18), note that if (12.1.19) has a solution $u_n \in V$, then

$$u_n = \frac{1}{\lambda}\,(P_n f + P_n K u_n) \in V_n.$$

Thus $P_n u_n = u_n$,

$$(\lambda - P_n K)\, u_n = P_n(\lambda - K)\, u_n,$$

and this shows that (12.1.19) implies (12.1.18).

For the error analysis, we compare (12.1.19) with the original equation $(\lambda - K)\, u = f$ of (12.1.3), since both equations are defined on the original space V. The theoretical analysis is based on the approximation of $\lambda - P_n K$ by $\lambda - K$:

$$\begin{aligned} \lambda - P_n K &= (\lambda - K) + (K - P_n K) \\ &= (\lambda - K)[I + (\lambda - K)^{-1}(K - P_n K)]. \end{aligned} \tag{12.1.20}$$

We use this in the following theorem.

Theorem 12.1.2 *Assume $K : V \to V$ is bounded, with V a Banach space; and assume $\lambda - K : V \underset{onto}{\overset{1-1}{\to}} V$. Further assume*

$$\|K - P_n K\| \to 0 \quad \text{as } n \to \infty. \tag{12.1.21}$$

Then for all sufficiently large n, say $n \geq N$, the operator $(\lambda - P_n K)^{-1}$ exists as a bounded operator from V to V. Moreover, it is uniformly bounded:

$$\sup_{n \geq N} \left\|(\lambda - P_n K)^{-1}\right\| < \infty. \tag{12.1.22}$$

For the solutions u_n with n sufficiently large and u of (12.1.19) *and* (12.1.3)*, respectively, we have*

$$u - u_n = \lambda(\lambda - P_n K)^{-1}(u - P_n u) \tag{12.1.23}$$

and the two-sided error estimate

$$\frac{|\lambda|}{\|\lambda - P_n K\|} \|u - P_n u\| \leq \|u - u_n\| \leq |\lambda| \left\|(\lambda - P_n K)^{-1}\right\| \|u - P_n u\|. \tag{12.1.24}$$

This leads to a conclusion that $\|u - u_n\|$ converges to zero at exactly the same speed as $\|u - P_n u\|$.

Proof. (a) Pick an N such that

$$\epsilon_N \equiv \sup_{n \geq N} \|K - P_n K\| < \frac{1}{\|(\lambda - K)^{-1}\|}. \tag{12.1.25}$$

Then the inverse $[I + (\lambda - K)^{-1}(K - P_n K)]^{-1}$ exists and is uniformly bounded by the geometric series theorem (Theorem 2.3.1 in Chapter 2), and

$$\left\|[I + (\lambda - K)^{-1}(K - P_n K)]^{-1}\right\| \leq \frac{1}{1 - \epsilon_N \|(\lambda - K)^{-1}\|}.$$

Using (12.1.20), we see that $(\lambda - P_n K)^{-1}$ exists,

$$(\lambda - P_n K)^{-1} = [I + (\lambda - K)^{-1}(K - P_n K)]^{-1}(\lambda - K)^{-1},$$

and then

$$\left\|(\lambda - P_n K)^{-1}\right\| \leq \frac{\left\|(\lambda - K)^{-1}\right\|}{1 - \epsilon_N \|(\lambda - K)^{-1}\|} \equiv M. \tag{12.1.26}$$

This shows (12.1.22).

(b) For the error formula (12.1.23), apply P_n to the equation $(\lambda - K)\, u = f$, and then rearrange to obtain

$$(\lambda - P_n K)\, u = P_n f + \lambda(u - P_n u).$$

Subtract $(\lambda - P_n K)\, u_n = P_n f$ from the above equality to get

$$(\lambda - P_n K)\,(u - u_n) = \lambda(u - P_n u). \tag{12.1.27}$$

Then

$$u - u_n = \lambda(\lambda - P_n K)^{-1}(u - P_n u),$$

which is (12.1.23). Taking norms and using (12.1.26),

$$\|u - u_n\| \leq |\lambda|\, M\, \|u - P_n u\|. \tag{12.1.28}$$

Thus if $P_n u \to u$, then $u_n \to u$ as $n \to \infty$.
(c) The upper bound in (12.1.24) follows directly from (12.1.23). The lower bound follows by taking norms in (12.1.27),

$$|\lambda|\, \|u - P_n u\| \leq \|\lambda - P_n K\|\, \|u - u_n\|\,.$$

This is equivalent to the lower bound in (12.1.24).

To obtain a lower bound which is uniform in n, note that for $n \geq N$,

$$\begin{aligned} \|\lambda - P_n K\| &\leq \|\lambda - K\| + \|K - P_n K\| \\ &\leq \|\lambda - K\| + \epsilon_N. \end{aligned}$$

The lower bound in (12.1.24) can now be replaced by

$$\frac{|\lambda|}{\|\lambda - K\| + \epsilon_N}\, \|u - P_n u\| \leq \|u - u_n\|.$$

Combining this inequality with (12.1.28), we have

$$\frac{|\lambda|}{\|\lambda - K\| + \epsilon_N}\, \|u - P_n u\| \leq \|u - u_n\| \leq |\lambda|\, M\, \|u - P_n u\|. \tag{12.1.29}$$

This shows that u_n converges to u if and only if $P_n u$ converges to u. Moreover, if convergence does occur, then $\|u - P_n u\|$ and $\|u - u_n\|$ tend to zero with exactly the same speed. □

We note that in order for the theorem to be true, it is necessary only that (12.1.25) be valid, not the stronger assumption of (12.1.21). Nonetheless, the theorem is applied usually by proving (12.1.21). Therefore, to apply the above theorem we need to know whether $\|K - P_n K\| \to 0$ as $n \to \infty$. The following two lemmas address this question.

Lemma 12.1.3 *Let V, W be Banach spaces, and let $A_n : V \to W$, $n \geq 1$, be a sequence of bounded linear operators. Assume $\{A_n u\}$ converges for all $u \in V$. Then the convergence is uniform on compact subsets of V.*

Proof. By the principle of uniform boundedness (Theorem 2.4.4 in Chapter 2), the operators A_n are uniformly bounded:

$$M \equiv \sup_{n \geq 1} \|A_n\| < \infty.$$

The operators A_n are also equicontinuous:

$$\|A_n u - A_n f\| \leq M \|u - f\|.$$

Let S be a compact subset of V. Then $\{A_n\}$ is a uniformly bounded and equicontinous family of functions on the compact set S; and it is then a standard result of analysis (a straightforward generalization of Ascoli's Theorem 1.6.3 in the setting of Banach spaces) that $\{A_n u\}$ is uniformly convergent for $u \in S$. □

Lemma 12.1.4 *Let V be a Banach space, and let $\{P_n\}$ be a family of bounded projections on V with*

$$P_n u \to u \quad \text{as } n \to \infty, \quad u \in V. \tag{12.1.30}$$

If $K : V \to V$ is compact, then

$$\|K - P_n K\| \to 0 \quad \text{as } n \to \infty.$$

Proof. From the definition of operator norm,

$$\|K - P_n K\| = \sup_{\|u\| \leq 1} \|Ku - P_n K u\| = \sup_{z \in K(U)} \|z - P_n z\|,$$

with $K(U) = \{Ku \mid \|u\| \leq 1\}$. The set $\overline{K(U)}$ is compact. Therefore, by the preceding Lemma 12.1.3 and the assumption (12.1.30),

$$\sup_{z \in K(U)} \|z - P_n z\| \to 0 \quad \text{as } n \to \infty.$$

This proves the lemma. □

This last lemma includes most cases of interest, but not all. There are situations where $P_n u \to u$ for most $u \in V$, but not all u. In such cases, it is necessary to show directly that $\|K - P_n K\| \to 0$, if it is true. Of course, in such cases, we see from (12.1.24) that $u_n \to u$ if and only if $P_n u \to u$; and thus the method is not convergent for some solutions u. This would occur, for example, if V_n is the set of polynomials of degree $\leq n$ and $V = C[a, b]$.

Exercise 12.1.1 Show that the condition (12.1.6) implies the set of functions $\{\phi_1, \ldots, \phi_\kappa\}$ is linearly independent over D.

Exercise 12.1.2 Prove the result stated in the last sentence of the proof of Lemma 12.1.3.

Exercise 12.1.3 Prove that the upper bound in (12.1.24) can be replaced by

$$\|u - u_n\| \leq |\lambda| \, (1 + \gamma_n) \left\|(\lambda - K)^{-1}\right\| \|u - P_n u\|$$

with $\gamma_n \to 0$ as $n \to \infty$.

Exercise 12.1.4 In Theorem 12.1.2, write

$$u - u_n = e_n^{(1)} + e_n^{(2)}, \tag{12.1.31}$$

with

$$e_n^{(1)} = \lambda(\lambda - K)^{-1}(u - P_n u)$$

and $e_n^{(2)}$ defined implicitly by this and (12.1.31). Show that under the assumptions of Theorem 12.1.2,

$$\|e_n^{(2)}\| \leq \delta_n \|e_n^{(1)}\|$$

with $\delta_n \to 0$ as $n \to \infty$.

Exercise 12.1.5 Let V and W be Banach spaces, let $A : V \xrightarrow[onto]{1-1} W$ be bounded, and let $B : V \to W$ be compact. Consider solving the equation $(A + B)\, u = f$ with the assumption that

$$(A + B)\, v = 0 \quad \Longrightarrow \quad v = 0.$$

Let V_n be an approximating finite dimensional subspace of V, and further assume V_n is a subspace of W. Let $P_n : V \to V_n$ be a bounded projection for which

$$P_n v \to v \quad \text{as } n \to \infty,$$

for all $v \in V$. Assume further that

$$P_n A = A P_n.$$

Consider the numerical approximation

$$P_n \, (A + B)\, u_n = P_n f, \quad u_n \in V_n.$$

Develop a stability and convergence analysis for this numerical method.
Hint: Consider the equation $\left(I + A^{-1}B\right) u = A^{-1} f$.

Exercise 12.1.6 Assuming $(\lambda - \mathcal{K})\, u = f$ is uniquely solvable in $L^2(D)$, consider the following numerical method for solving it. Let $\mathcal{X}_n$ be the span of the linearly independent functions $\phi_1, \ldots, \phi_\kappa$, with $\kappa \equiv \kappa_n$. Determine $u_n \in \mathcal{X}_n$ by choosing it as the minimizer of $\|\, (\lambda - \mathcal{K})\, v - f\|_{L^2}$ as v ranges over $\mathcal{X}_n$. This is called the 'least squares method' for the approximate solution of $(\lambda - \mathcal{K})\, u = f$. Introduce the new subspace

$$\mathcal{Y}_n = \{(\lambda - \mathcal{K})\, v \mid v \in \mathcal{X}_n\}$$

and let $\mathcal{P}_n$ be the orthogonal projection of $L^2(D)$ onto $\mathcal{Y}_n$. Show that the minimizer u_n is the solution of the projection equation (12.1.19).
Hint: Write u_n in terms of the basis $\{\phi_1, \dots, \phi_\kappa\}$:

$$u_n = \sum_{j=1}^{\kappa_n} c_j \phi_j.$$

Seek to minimize

$$\| (\lambda - \mathcal{K}) u_n - f \|_{L^2}^2 = ((\lambda - \mathcal{K}) u_n - f, (\lambda - \mathcal{K}) u_n - f),$$

considered as a function of the coefficients $c_1, \dots, c_\kappa$.

12.2 Examples

Most projection methods are based on the ways in which we approximate functions, and there are two main approaches.

- Decompose the approximation region D into elements $\Delta_1, \dots, \Delta_m$; and then approximate a function $u \in C(D)$ by a low degree polynomial over each of the elements Δ_i. These projection methods are often referred to as *piecewise polynomial methods* or *finite element methods*; and when D is the boundary of a region, such methods are often called *boundary element methods*.
- Approximate $u \in C(D)$ by using a family of functions which are defined globally over all of D, for example, use polynomials, trigonometric polynomials, or spherical polynomials. Often, these approximating functions are also infinitely differentiable. Sometimes these types of projection methods are referred to as *spectral methods*, especially when trigonometric polynomials are used.

We illustrate each of these, relying on approximation results introduced earlier in Chapter 3.

12.2.1 Piecewise linear collocation

We consider the numerical solution of the integral equation

$$\lambda u(x) - \int_a^b k(x,y) u(y)\, dy = f(x), \quad a \le x \le b, \tag{12.2.1}$$

using piecewise linear approximating functions. Recall the definition of piecewise linear interpolation given in Subsection 3.2.3, including the piecewise linear interpolatory projection operator of (3.2.7). For convenience, we repeat those results here. Let $D = [a, b]$ and $n \ge 1$, and define $h = (b-a)/n$,

$$x_j = a + jh, \quad j = 0, 1, \dots, n.$$

The subspace V_n is the set of all functions which are continuous and piecewise linear on $[a, b]$, with breakpoints $\{x_0, \ldots, x_n\}$. Its dimension is $n + 1$.

Introduce the Lagrange basis functions for continuous piecewise linear interpolation:

$$\ell_i(x) = \begin{cases} 1 - \dfrac{|x - x_i|}{h}, & x_{i-1} \le x \le x_{i+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{12.2.2}$$

with the obvious adjustment of the definition for $\ell_0(x)$ and $\ell_n(x)$. The projection operator is defined by

$$P_n u(x) = \sum_{i=0}^{n} u(x_i)\ell_i(x). \tag{12.2.3}$$

For convergence of $P_n u$, recall from (3.2.8) and (3.2.9) that

$$\|u - P_n u\|_\infty \le \begin{cases} \omega(u, h), & u \in C[a, b], \\ \dfrac{h^2}{8} \|u''\|_\infty, & u \in C^2[a, b]. \end{cases} \tag{12.2.4}$$

This shows that $P_n u \to u$ for all $u \in C[a, b]$; and for $u \in C^2[a, b]$, the convergence order is 2. For any compact operator $K : C[a, b] \to C[a, b]$, Lemma 12.1.4 implies $\|K - P_n K\| \to 0$ as $n \to \infty$. Therefore the results of Theorem 12.1.2 can be applied directly to the numerical solution of the integral equation $(\lambda - K)\, u = f$. For sufficiently large n, say $n \ge N$, the equation $(\lambda - P_n K)\, u_n = P_n f$ has a unique solution u_n for each $f \in C[a, b]$. Assuming $u \in C^2[a, b]$, (12.1.24) implies

$$\|u - u_n\|_\infty \le |\lambda|\, M\, \frac{h^2}{8} \|u''\|_\infty, \tag{12.2.5}$$

with M a uniform bound on $(\lambda - P_n K)^{-1}$ for $n \ge N$.

The linear system (12.1.5) takes the simpler form

$$\lambda u_n(x_i) - \sum_{j=0}^{n} u_n(x_j) \int_a^b k(x_i, y)\ell_j(y)\, dy = f(x_i), \quad i = 0, \ldots, n. \tag{12.2.6}$$

The integrals can be simplified. For $j = 1, \ldots, n - 1$,

$$\begin{aligned} \int_a^b k(x_i, y)\ell_j(y)\, dy = {} & \frac{1}{h} \int_{x_{j-1}}^{x_j} k(x_i, y)(y - x_{j-1})\, dy \\ & + \frac{1}{h} \int_{x_j}^{x_{j+1}} k(x_i, y)(x_j - y)\, dy. \end{aligned} \tag{12.2.7}$$

The integrals for $j = 0$ and $j = n$ are modified straightforwardly. These integrals must usually be calculated numerically; and we want to use a quadrature method which retains the order of convergence in (12.2.5) at a minimal cost in calculation time.

n	$E_n^{(1)}$	*Ratio*	$E_n^{(2)}$	*Ratio*
2	$5.25E-3$		$2.32E-2$	
4	$1.31E-3$	4.01	$7.91E-3$	2.93
8	$3.27E-4$	4.01	$2.75E-3$	2.88
16	$8.18E-5$	4.00	$9.65E-4$	2.85
32	$2.04E-5$	4.00	$3.40E-4$	2.84
64	$5.11E-6$	4.00	$1.20E-4$	2.83
128	$1.28E-6$	4.00	$4.24E-5$	2.83

TABLE 12.1. Example of piecewise linear collocation for solving (12.2.8)

Example 12.2.1 Consider the integral equation

$$\lambda u(x) - \int_0^b e^{xy} u(y)\, dy = f(x), \quad 0 \le x \le b. \tag{12.2.8}$$

The equation parameters are $b = 1,\ \lambda = 5$. We use the two unknowns

$$u^{(1)}(x) = e^{-x}\cos(x), \quad u^{(2)}(x) = \sqrt{x}, \quad 0 \le x \le b \tag{12.2.9}$$

and define $f(x)$ accordingly. The results of the use of piecewise linear collocation are given in Table 12.1. The errors given in the table are the maximum errors on the collocation node points,

$$E_n^{(k)} = \max_{0 \le i \le n} |u^{(k)}(x_i) - u_n^{(k)}(x_i)|.$$

The column labeled *Ratio* is the ratio of the successive values of $E_n^{(k)}$ as n is doubled.

The function $u^{(2)}(x)$ is not continuously differentiable on $[0, b]$, and we have no reason to expect a rate of convergence of $\mathcal{O}(h^2)$. Empirically, the errors $E_n^{(2)}$ appear to be $\mathcal{O}(h^{3/2})$. From (12.1.24), Theorem 12.1.2, we know that $\|u^{(2)} - u_n^{(2)}\|_\infty$ converges to zero at exactly the same speed as $\|u^{(2)} - P_n u^{(2)}\|_\infty$, and it can be shown that the latter is only $\mathcal{O}(h^{1/2})$. This apparent contradiction between the empirical and theoretical rates is due to $u_n(t)$ being *superconvergent* at the collocation node points: for the numerical solution $u_n^{(2)}$,

$$\lim_{n\to\infty} \frac{E_n^{(2)}}{\|u^{(2)} - u_n^{(2)}\|_\infty} = 0.$$

This is examined in much greater detail in the following Section 12.3. □

12.2.2 Trigonometric polynomial collocation

We solve the integral equation

$$\lambda u(x) - \int_0^{2\pi} k(x,y)u(y)\,dy = f(x), \quad 0 \le x \le 2\pi, \tag{12.2.10}$$

in which the kernel function is assumed to be continuous and 2π-periodic in both y and x:

$$k(x+2\pi, y) \equiv k(x, y+2\pi) \equiv k(x,y).$$

Let $V = C_p(2\pi)$, space of all 2π-periodic and continuous functions on $\mathbb{R}$. We consider the solution of (12.2.10) for $f \in C_p(2\pi)$, which then implies $u \in C_p(2\pi)$.

Since the solution $u(x)$ is 2π-periodic, we approximate it with *trigonometric polynomials;* and we use the general framework for trigonometric polynomial interpolation of Subsection 3.7.3 from Chapter 3. Let V_n denote the trigonometric polynomials of degree at most n; and recall V_n has dimension $\kappa_n = 2n+1$. Let $\{\phi_1(x), \dots, \phi_\kappa(x)\}$ denote a basis for V_n, either $\left\{e^{ikx} \mid k = 0, \pm 1, \dots, \pm n\right\}$ or

$$\{1, \sin x, \cos x, \dots, \sin nx, \cos nx\}. \tag{12.2.11}$$

The interpolatory projection of $C_p(2\pi)$ onto V_n is given by

$$P_n u(x) = \sum_{j=1}^{\kappa_n} u(x_j)\ell_j(x), \tag{12.2.12}$$

where the Lagrange basis functions $\ell_j(x)$ are given implicitly in the Lagrange formula (3.7.19) of Section 3.7.3. Note that P_n was denoted by $\mathcal{I}_n$ in that formula.

From (3.7.20), $\|P_n\| = \mathcal{O}(\log n)$. Since $\|P_n\| \to \infty$ as $n \to \infty$, it follows from the principle of uniform boundedness that there exists $u \in C_p(2\pi)$ for which $P_n u$ does not converge to u in $C_p(2\pi)$ (Theorem 2.4.4 in Chapter 2).

Consider the use of the above trigonometric interpolation in solving (12.2.10) by collocation. The linear system (12.1.5) becomes

$$\sum_{j=1}^{\kappa_n} c_j \left[\lambda\phi_j(x_i) - \int_0^{2\pi} k(x_i,y)\phi_j(y)\,dy\right] = f(x_i), \quad i = 1, \dots, \kappa_n, \tag{12.2.13}$$

and the solution is

$$u_n(x) = \sum_{j=1}^{\kappa_n} c_j \phi_j(x).$$

The integrals in (12.2.13) are usually evaluated numerically, and for that, we recommend using the trapezoidal rule. With periodic integrands, the

trapezoidal rule is very effective, as was noted earlier in Proposition 7.5.6 of Chapter 7.

To prove the convergence of this collocation method, we must show

$$\|K - P_n K\| \to 0 \quad \text{as } n \to \infty.$$

Since Lemma 12.1.4 cannot be used, we must examine $\|K - P_n K\|$ directly. The operator $P_n K$ is an integral operator, with

$$\begin{aligned} P_n K u(x) &= \int_0^{2\pi} k_n(x,y) u(y)\, dy, \\ k_n(x,y) &\equiv (P_n k_y)(x), \quad k_y(x) \equiv k(x,y). \end{aligned} \tag{12.2.14}$$

To show convergence of $\|K - P_n K\|$ to zero, we must prove directly that

$$\|K - P_n K\| = \max_{0 \le x \le 2\pi} \int_0^{2\pi} |k(x,y) - k_n(x,y)|\, dy \tag{12.2.15}$$

converges to zero. To do so, we use the result (3.7.21) on the convergence of trigonometric polynomial interpolation.

Assume that $k(x,y)$ satisfies, for some $\alpha > 0$,

$$|k(x,y) - k(\xi,y)| \le c(k)\, |x - \xi|^{\alpha}, \tag{12.2.16}$$

for all y, x, ξ. Then we leave it as Exercise 12.2.2 to prove that

$$\|K - P_n K\| \le \frac{c \log n}{n^{\alpha}}. \tag{12.2.17}$$

Since this converges to zero, we can apply Theorem 12.1.2 to the error analysis of the collocation method with trigonometric interpolation.

Assuming (12.2.10) is uniquely solvable, the collocation equation

$$(\lambda - P_n K)\, u_n = P_n f$$

has a unique solution u_n for all sufficiently large n; and $\|u - u_n\|_\infty \to 0$ if and only if $\|u - P_n u\|_\infty \to 0$. We know there are cases for which the latter is not true; but from (3.7.21) of Section 3.7 in Chapter 3, $\|u - u_n\|_\infty \nrightarrow 0$ only for functions u with very little smoothness (Theorem 3.7.1 with $k = 0$). For functions u which are infinitely differentiable, the bound (3.7.21) shows the rate of convergence is very rapid, faster than $\mathcal{O}(n^{-k})$ for any k.

There are kernel functions $k(x,y)$ which do not satisfy (12.2.16), but to which the above collocation method can still be applied. Their error analysis requires a more detailed knowledge of the smoothing properties of the operator K. Such cases occur when solving boundary integral equations with singular kernel functions, such as that defined in (7.5.16) of Section 7.5.

12.2.3 A piecewise linear Galerkin method

The error analysis of Galerkin methods is usually carried out in a Hilbert space, generally $L^2(D)$ or some Sobolev space $H^r(D)$. Following this, an analysis within $C(D)$ is often also given, to obtain results on uniform convergence of the numerical solutions.

We again consider the numerical solution of (12.2.1). Let $V = L^2(a,b)$, and let its norm and inner product be denoted by simply $\|\cdot\|$ and $(\cdot,\cdot)$, respectively. Let V_n be the subspace of continuous piecewise linear functions as described earlier in Subsection 12.2.1. The dimension of V_n is $n+1$, and the Lagrange functions of (12.2.2) are a basis for V_n. However, now P_n denotes the orthogonal projection of $L^2(a,b)$ onto V_n. We begin by showing that $P_n u \to u$ for all $u \in L^2(a,b)$.

Begin by assuming $u(x)$ is continuous on $[a,b]$. Let $\mathcal{I}_n u(x)$ denote the piecewise linear function in V_n which interpolates $u(x)$ at $x = x_0, \ldots, x_n$; see (12.2.3). Recall that $P_n u$ minimizes $\|u - z\|$ as z ranges over V_n, a fact which is expressed in the identity in Proposition 3.6.9(c). Therefore,

$$\begin{aligned} \|u - P_n u\| &\le \|u - \mathcal{I}_n u\| \\ &\le \sqrt{b-a}\, \|u - \mathcal{I}_n u\|_\infty \\ &\le \sqrt{b-a}\, \omega(u;h). \end{aligned} \tag{12.2.18}$$

The last inequality uses the error bound (12.2.4). This shows $P_n u \to u$ for all continuous functions u on $[a,b]$.

It is well known that the set of all continuous functions on $[a,b]$ is dense in $L^2(a,b)$ (see the comment following Theorem 1.5.6). Also, the orthogonal projection P_n satisfies $\|P_n\| = 1$; see (3.4.5) of Section 3.4. For a given $u \in L^2(a,b)$, let $\{u_m\}$ be a sequence of continuous functions which converge to u in $L^2(a,b)$. Then

$$\begin{aligned} \|u - P_n u\| &\le \|u - u_m\| + \|u_m - P_n u_m\| + \|P_n(u - u_m)\| \\ &\le 2\,\|u - u_m\| + \|u_m - P_n u_m\|. \end{aligned}$$

Given an $\epsilon > 0$, pick m such that $\|u - u_m\| < \epsilon/4$; and fix m. This then implies that for all n,

$$\|u - P_n u\| \le \frac{\epsilon}{2} + \|u_m - P_n u_m\|.$$

We have

$$\|u - P_n u\| \le \epsilon$$

for all sufficiently large values of n. Since ϵ was arbitrary, this shows that $P_n u \to u$ for general $u \in L^2(a,b)$.

For the integral equation $(\lambda - K)\, u = f$, we can use Lemma 12.1.4 to obtain $\|K - P_n K\| \to 0$. This justifies the use of Theorem 12.1.2 to carry out the error analysis for the Galerkin equation $(\lambda - P_n K)\, u_n = P_n f$. As

before, $\|u - u_n\|$ converges to zero with the same speed as $\|u - P_n u\|$. For $u \in C^2[a,b]$, we combine (12.1.28), (12.2.18), and (12.2.4), to obtain

$$\begin{aligned}\|u - u_n\| &\leq |\lambda|\, M\, \|u - P_n u\| \\ &\leq |\lambda|\, M\sqrt{b-a}\, \|u - \mathcal{I}_n u\|_\infty \\ &\leq |\lambda|\, M\sqrt{b-a}\, \frac{h^2}{8} \|u''\|_\infty. \end{aligned} \tag{12.2.19}$$

For the linear system, we use the Lagrange basis functions of (12.2.2). These are not orthogonal, but they are still a very convenient basis with which to work. Moreover, producing an orthogonal basis for V_n is a non-trivial task. The solution u_n of $(\lambda - P_n K)u_n = P_n f$ is given by

$$u_n(x) = \sum_{j=0}^{n} c_j \ell_j(x).$$

The coefficients $\{c_j\}$ are obtained by solving the linear system

$$\begin{aligned}\sum_{j=0}^{n} c_j \left[\lambda(\ell_i, \ell_j) - \int_a^b \int_a^b k(x,y)\ell_i(x)\ell_j(y)\, dy\, dx\right] \\ = \int_a^b f(x)\ell_i(x)\, dx, \quad i = 0, \ldots, n. \end{aligned} \tag{12.2.20}$$

For the coefficients (ℓ_i, ℓ_j),

$$(\ell_i, \ell_j) = \begin{cases} 0, & |i-j| > 1, \\ \dfrac{2h}{3}, & 0 < i = j < n, \\ \dfrac{h}{3}, & i = j = 0 \text{ or } n, \\ \dfrac{h}{6}, & |i-j| = 1. \end{cases} \tag{12.2.21}$$

The double integrals in (12.2.20) reduce to integrals over much smaller subintervals, because the basis functions $\ell_i(x)$ are zero over most of $[a,b]$. If these integrals are evaluated numerically, it is important to evaluate them with an accuracy consistent with the error bound in (12.2.19). Lesser accuracy degrades the accuracy of the Galerkin solution u_n; and greater accuracy is an unnecessary expenditure of effort.

Just as was true with collocation methods, we can easily generalize the above presentation to include the use of piecewise polynomial functions of any fixed degree. Since the theory is entirely analogous to that presented above, we omit it here and leave it as an exercise for the reader.

We defer to the following case a consideration of the uniform convergence of $u_n(x)$ to $u(x)$.

12.2.4 A Galerkin method with trigonometric polynomials

We consider again the use of trigonometric polynomials as approximations in solving the integral equation (12.2.10), with $k(x,y)$ and $f(x)$ being 2π-periodic functions as before. Initially, we use the space $V = L^2(0,2\pi)$, the space of all complex-valued and square integrable Lebesgue measurable functions on $(0,2\pi)$. The inner product is defined by

$$(u,v) = \int_0^{2\pi} u(x)\overline{v(x)}\,dx.$$

Later we consider the space $C_p(2\pi)$, the set of all complex-valued 2π-periodic continuous functions, with the uniform norm.

The approximating subspace V_n is again the set of all trigonometric polynomials of degree $\leq n$. As a basis, we use the complex exponentials,

$$\phi_j(x) = e^{ijx}, \quad j = 0, \pm 1, \dots, \pm n.$$

Earlier, in Example 3.4.9 of Section 3.4 and in Subsection 3.7.1, we introduced and discussed the Fourier series

$$u(x) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} (u,\phi_j)\phi_j(x), \quad u \in L^2(0,2\pi), \tag{12.2.22}$$

with respect to the basis

$$\{1, \sin x, \cos x, \dots, \sin nx, \cos nx, \dots\}.$$

The basis

$$\left\{e^{ijx} \mid j = 0, \pm 1, \pm 2, \dots\right\}$$

was used in defining the periodic Sobolev spaces $H^r(0,2\pi)$ in Section 7.5 of Chapter 7. These two bases are equivalent, and it is straightforward to convert between them. It is well known that for $u \in L^2(0,2\pi)$, the Fourier series converges in the norm of $L^2(0,2\pi)$.

The orthogonal projection of $L^2(0,2\pi)$ onto V_n is just the n^{th} partial sum of this series,

$$P_n u(x) = \frac{1}{2\pi} \sum_{j=-n}^{n} (u,\phi_j)\phi_j(x), \tag{12.2.23}$$

which was denoted earlier by $\mathcal{F}_n u$ in Example 3.6.8 of Section 3.6. From the convergence of (12.2.22), it follows that $P_n u \to u$ for all $u \in L^2(0,2\pi)$. Its rate of uniform convergence was considered in Subsection 3.7.1. For its rate of convergence in $H^r(0,2\pi)$, it is straightforward to use the framework of Section 7.5 in Chapter 7 to prove the following:

$$\|u - P_n u\|_{L^2} \leq \frac{c}{n^r} \left[\frac{1}{2\pi} \sum_{|j|>n} |j|^{2r} |(u,\phi_j)|^2\right]^{\frac{1}{2}} \leq \frac{c}{n^r}\|u\|_{H^r} \tag{12.2.24}$$

for $u \in H^r(0, 2\pi)$.

Using Lemma 12.1.4, we have that $\|K - P_n K\| \to 0$ as $n \to \infty$. Thus Theorem 12.1.2 can be applied to the error analysis of the approximating equation $(\lambda - P_n K)\, u_n = P_n f$. For all sufficiently large n, say $n \geq N$, the inverses $(\lambda - P_n K)^{-1}$ are uniformly bounded; and $\|u - u_n\|$ can be bounded proportional to $\|u - P_n u\|$, and thus u_n converges to u. One result on the rate of convergence is obtained by applying (12.2.24) to (12.1.28):

$$\|u - u_n\| \leq \frac{c\,|\lambda|\,M}{n^r} \|u\|_{H^r}, \quad n \geq N,\ u \in H^r(0, 2\pi) \tag{12.2.25}$$

with c the same as in (12.2.24).

With respect to the basis $\{e^{ijx} \mid j = 0, \pm 1, \pm 2, \dots\}$, the linear system (12.1.14) for

$$(\lambda - P_n K) u_n = P_n f$$

is given by

$$\begin{aligned} 2\pi\lambda c_k - \sum_{j=-n}^{n} c_j \int_0^{2\pi} \int_0^{2\pi} e^{i(jy-kx)} k(x, y)\, dy\, dx \\ = \int_0^{2\pi} e^{-ikx} f(x)\, dx, \quad k = -n, \dots, n \end{aligned} \tag{12.2.26}$$

with the solution u_n given by

$$u_n(x) = \sum_{j=-n}^{n} c_j e^{ijx}.$$

The integrals in this system are usually evaluated numerically; and this is examined in some detail in [18, pp. 148–150]. Again, the trapezoidal rule is the standard form of quadrature used in evaluating these integrals; the fast Fourier transform can also be used to improve the efficiency of the quadrature process (see e.g. [15, p. 181], [120, Chap. 13]).

Another important example of the use of globally defined and smooth approximations is the use of spherical polynomials (see (7.5.5) in Chapter 7) as approximations to functions defined on the unit sphere in $\mathbb{R}^3$.

Uniform convergence

We often are interested in obtaining uniform convergence of u_n to u. For this, we regard the operator P_n of (12.2.23) as an operator on $C_p(2\pi)$ to V_n, and we take $V = C_p(2\pi)$. Unfortunately, it is no longer true that $P_n u$ converges to u for all $u \in V$, and consequently, Lemma 12.1.4 cannot be applied. In fact, from (3.7.9)–(3.7.10) of Subsection 3.7.1,

$$\|P_n\| = \mathcal{O}(\log n), \tag{12.2.27}$$

which implies the sequence $\{P_n\}$ is not uniformly bounded and therefore we do not expect pointwise convergence of the sequence $\{P_n u\}$ to u for all $u \in V$.

We use the framework of (12.2.14)–(12.2.15) to examine whether the quantity $\|K - P_n K\|$ converges to zero or not. In the present case, the projection used in (12.2.14) is the orthogonal Fourier projection of (12.2.23), but otherwise the results are the same.

Assume $k(x, y)$ satisfies the Hölder condition

$$|k(x, y) - k(\xi, y)| \leq c(K) |x - \xi|^{\alpha},$$

for all y, x, ξ, for some $0 < \alpha \leq 1$. Then apply (3.7.12) to obtain

$$\|K - P_n K\| \leq \frac{c \log n}{n^{\alpha}}$$

for a suitable constant c. With this, we can apply Theorem 12.1.2 and obtain a complete convergence analysis within $C_p(2\pi)$, thus obtaining results on the uniform convergence of u_n to u.

Another way of obtaining such uniform convergence results can be based on the ideas of the following section on the iterated projection method.

Conditioning of the linear system

We have omitted any discussion of the conditioning of the linear systems associated with either the Galerkin or collocation methods. This is important when implementing these methods, and the basis for V_n should be chosen with some care. The linear system is as well-conditioned as can be expected, based on the given equation $(\lambda - K) u = f$, if the matrix $(\phi_j(x_i)) = I$ for collocation or $((\phi_j, \phi_i)) = I$ for the Galerkin method. It can still be well-conditioned without such a restriction, but the choice of basis must be examined more carefully. See [18, Section 3.6] for an extended discussion.

Exercise 12.2.1 For the piecewise linear interpolatory projection operator of Subsection 12.2.1, calculate an explicit formula for the operator $P_n K$, showing it is a degenerate kernel integral operator. Be as explicit as possible in defining the degenerate kernel. Assuming $k(x, y)$ is twice continuously differentiable with respect to x, uniformly for $a \leq y \leq b$, show

$$\|K - P_n K\| \leq \frac{h^2}{8} \max_{a \leq x \leq b} \int_a^b \left| \frac{\partial^2 k(x, y)}{\partial x^2} \right| dy.$$

Exercise 12.2.2 Prove (12.2.17).
Hint: Apply Theorem 3.7.1.

Exercise 12.2.3 Generalize the ideas of Subsection 12.2.1 to continuous piecewise polynomial collocation of degree $\kappa > 0$.

Exercise 12.2.4 Generalize the ideas of Subsection 12.2.3 to V_n the set of piecewise linear functions in which there is no restriction that the approximations be continuous.

Exercise 12.2.5 Generalize the ideas of Subsection 12.2.3 to V_n being the set of piecewise linear functions in which there is no restriction that the approximations be continuous. What is the dimension of V_n? Show that the standard orthogonal Legendre polynomials defined on $[-1, 1]$ can be used to create an orthogonal basis for V_n.
Hint: Produce a local orthogonal basis on each subinterval $[x_{j-1}, x_j]$.

Exercise 12.2.6 Prove (12.2.24).

Exercise 12.2.7 Give conditions on the data λ, k, and f so that a solution of the equation (12.2.1) has the regularity $u \in C^2[a, b]$. Note that this regularity is required in the error estimate (12.2.5).

Exercise 12.2.8 Let P_n be an interpolatory projection operator, and let

$$Kv(x) = \int_a^b k(x, y)v(y)\, dy, \quad a \le x \le b,\ v \in C[a, b]$$

have a continuous kernel function $k(x, y)$. Show that P_nK is a degenerate kernel integral operator. For the case of P_n the piecewise linear interpolatory operator of (12.2.3), write out an explicit formula for the degenerate kernel $k_n(x, y)$ and analyze the error $k(x, y) - k_n(x, y)$.

Exercise 12.2.9 It is known that if $u \in C[-1, 1]$, then the partial sums of the Chebyshev expansion

$$u(x) = \frac{c_0}{2} + \sum_{i=1}^{\infty} c_i T_i(x), \quad c_i = \frac{2}{\pi} \int_{-1}^{1} \frac{u(y)T_i(y)}{\sqrt{1 - y^2}}\, dy$$

are good uniform approximations of $u(x)$ when u is sufficiently smooth. This is an orthogonal polynomial expansion of u. The weight function is $w(y) = 1/\sqrt{1 - y^2}$, and the associated orthogonal family is the Chebyshev polynomials $\{T_i(x)\}_{i\ge 0}$. We want to investigate the solution of

$$\lambda u(x) - \int_{-1}^{1} k(x, y)u(y)\, dy = f(x), \quad -1 \le x \le 1$$

using Galerkin's method with polynomial subspaces and the orthogonal projections

$$P_n v(x) = \frac{c_0}{2} + \sum_{i=1}^{\infty} c_i T_i(x), \quad n \ge 1.$$

The space being used is $L^2_w(-1, 1)$ with the $w(y)$ given above.
(a) Give the Galerkin method for solving the above integral equation.
(b) Give the coefficients of the linear system, and suggest a way for dealing with the singularity in the integrand (due to the presence of the weight function w).

(c) If the true solution u is r-times continuously differentiable on $[-1,1]$, discuss the rate of convergence to zero of the error $\|u-u_n\|_{L^2_w}$. For an introductory account of Chebyshev polynomials and Chebyshev expansions, see [15, Sections 4.5–4.7].

Exercise 12.2.10 Recall the linear system (12.1.5) for the collocation method, and consider it with the Lagrange basis $\{\ell_i(x)\}$ satisfying (12.1.7), with the associated projection operator P_n of (12.1.8). Denote this linear system by $A_n\boldsymbol{u}_n = \boldsymbol{f}_n$, with

$$\boldsymbol{f}_n = (f(x_1),\dots,f(x_\kappa))^T$$

and $\boldsymbol{u}_n$ defined analogously. For A_n,

$$(A_n)_{i,j} = \lambda\delta_{i,j} - \int_D k(x_i,y)\ell_j(y)\,dy.$$

Consider $A_n : \mathbb{R}^\kappa \to \mathbb{R}^\kappa$ with the infinity norm, and find a bound for $\|A_n^{-1}\|$, using the row norm as the matrix norm. Find the bound in terms of $\|(\lambda - P_nK)^{-1}\|$. *Hint*: For arbitrary $\boldsymbol{b}\in\mathbb{R}^\kappa$, let $\boldsymbol{v} = A_n^{-1}\boldsymbol{b}$, or equivalently, $A_n\boldsymbol{v} = \boldsymbol{b}$. You need to bound $\boldsymbol{v}$ in terms of $\boldsymbol{b}$. To do this, begin by showing you can construct $g\in C(D)$ with

$$\boldsymbol{b} = (g(x_1),\dots,g(x_\kappa))^T$$

and $\|g\|_\infty = \|\boldsymbol{b}\|_\infty$. Define the function $v\in C(D)$ as the solution of

$$(\lambda - P_nK)\,v = P_ng.$$

Then bound $\boldsymbol{v}$ in terms of $\|v\|_\infty$, and bound the latter in terms of $\|\boldsymbol{b}\|_\infty$.

Exercise 12.2.11 To implement the piecewise linear collocation method given in (12.2.6), the integrals in the linear system must be approximated.

a. Show that if all elements of the matrix in (12.2.6) are approximated with an accuracy of $\mathcal{O}(h^3)$, then the resulting numerical solution $\widetilde{u}_n$ will be in error from u_n by $\mathcal{O}(h^2)$, thus preserving the accuracy associated with u_n as an approximation to u.

b. Give an approximation of (12.2.7) that is error by $\mathcal{O}(h^3)$.

12.3 Iterated projection methods

For the integral equation $(\lambda - K)\,u = f$, consider the following fixed point iteration which was considered earlier in (5.2.9) of Section 5.2, Chapter 5:

$$u^{(k+1)} = \frac{1}{\lambda}\left[f + Ku^{(k)}\right], \quad k = 0,1,\dots$$

As earlier, this iteration can be shown to converge to the solution u if $\|K\| < |\lambda|$; and in that case

$$\|u - u^{(k+1)}\| \le \frac{\|K\|}{|\lambda|}\left\|u - u^{(k)}\right\|.$$

In the paper [205], Sloan showed that one such iteration is always a good idea if K is a compact operator and if the initial guess is the solution u_n obtained by the Galerkin method, regardless of the size of $\|K\|$. We examine this idea and its consequences for projection methods.

Let u_n be the solution of the projection equation $(\lambda - P_n K)\, u_n = P_n f$. Define the *iterated projection solution* by

$$\widehat{u}_n = \frac{1}{\lambda}\,(f + K u_n)\,. \tag{12.3.1}$$

This new approximation $\widehat{u}_n$ is often an improvement upon u_n. Moreover, it can often be used to better understand the behaviour of the original projection solution u_n.

Applying P_n to both sides of (12.3.1), we have

$$P_n \widehat{u}_n = \frac{1}{\lambda}\,(P_n f + P_n K u_n)\,,$$

i.e.

$$P_n \widehat{u}_n = u_n. \tag{12.3.2}$$

Thus, u_n is the projection of $\widehat{u}_n$ into V_n. Substituting into (12.3.1) and rearranging terms, we have $\widehat{u}_n$ satisfies the equation

$$(\lambda - K P_n)\widehat{u}_n = f. \tag{12.3.3}$$

Often we can directly analyze this equation; and then information can be obtained on u_n by applying (12.3.2).

Also, since

$$u - \widehat{u}_n = \frac{1}{\lambda}\,(f + Ku) - \frac{1}{\lambda}\,(f + K u_n) = \frac{1}{\lambda} K(u - u_n), \tag{12.3.4}$$

we have the error bound

$$\|u - \widehat{u}_n\| \leq \frac{1}{|\lambda|}\|K\|\,\|u - u_n\|. \tag{12.3.5}$$

This proves the convergence of $\widehat{u}_n$ to u is at least as rapid as that of u_n to u. Often it is more rapid, because operating on $u - u_n$ with K, as in (12.3.4), sometimes causes cancellation due to the smoothing behaviour of integration.

From the above, we see that if $(\lambda - P_n K)^{-1}$ exists, then so does $(\lambda - K P_n)^{-1}$. Moreover, from the definition of the solution u_n and (12.3.1), we have

$$\widehat{u}_n = \frac{1}{\lambda}\,(f + K u_n) = \frac{1}{\lambda}\left[f + K(\lambda - P_n K)^{-1} P_n f\right];$$

and when combined with (12.3.3),

$$(\lambda - K P_n)^{-1} = \frac{1}{\lambda}\left[I + K(\lambda - P_n K)^{-1} P_n\right]. \tag{12.3.6}$$

Conversely, if $(\lambda - KP_n)^{-1}$ exists, then so does $(\lambda - P_nK)^{-1}$. This follows from the general lemma given below, which also shows that

$$(\lambda - P_nK)^{-1} = \frac{1}{\lambda}\left[I + P_n(\lambda - KP_n)^{-1}K\right]. \tag{12.3.7}$$

By combining (12.3.6) and (12.3.7), or by returning to the definitions of u_n and $\widehat{u}_n$, we also have

$$(\lambda - P_nK)^{-1}P_n = P_n(\lambda - KP_n)^{-1}. \tag{12.3.8}$$

We can choose to show the existence of either $(\lambda - P_nK)^{-1}$ or $(\lambda - KP_n)^{-1}$, whichever is the more convenient; and the existence of the other inverse follows immediately. Bounds on one inverse in terms of the other can also be derived by using (12.3.6) and (12.3.7).

Lemma 12.3.1 *Let V be a Banach space, and let A, B be bounded linear operators on V to V. Assume $(\lambda - AB)^{-1}$ exists from V onto V. Then $(\lambda - BA)^{-1}$ also exists, and*

$$(\lambda - BA)^{-1} = \frac{1}{\lambda}\left[I + B(\lambda - AB)^{-1}A\right]. \tag{12.3.9}$$

Proof. Calculate

$$\begin{aligned}
&(\lambda - BA)\frac{1}{\lambda}\left[I + B(\lambda - AB)^{-1}A\right] \\
&\quad = \frac{1}{\lambda}\left[\lambda - BA + (\lambda - BA)B(\lambda - AB)^{-1}A\right] \\
&\quad = \frac{1}{\lambda}\left[\lambda - BA + B(\lambda - AB)(\lambda - AB)^{-1}A\right] \\
&\quad = \frac{1}{\lambda}(\lambda - BA + BA) \\
&\quad = I.
\end{aligned}$$

A similar proof works to show

$$\frac{1}{\lambda}\left[I + B(\lambda - AB)^{-1}A\right](\lambda - BA) = I. \tag{12.3.10}$$

This proves (12.3.9). □

For the error in $\widehat{u}_n$, first rewrite $(\lambda - K)u = f$ as

$$(\lambda - KP_n)u = f + Ku - KP_nu.$$

Subtract (12.3.3) to obtain

$$(\lambda - KP_n)(u - \widehat{u}_n) = K(I - P_n)u. \tag{12.3.11}$$

Below we examine this apparently simple equation in much greater detail.

12.3.1 The iterated Galerkin method

Assume that V is a Hilbert space and that u_n is the Galerkin solution of the equation $(\lambda - K)u = f$ over a finite-dimensional subspace $V_n \subset V$. Then

$$(I - P_n)^2 = I - P_n.$$

and

$$\begin{aligned} \|K(I - P_n)u\| &= \|K(I - P_n)(I - P_n)u\| \\ &\leq \|K(I - P_n)\| \, \|(I - P_n)u\| . \end{aligned} \tag{12.3.12}$$

Using the fact that we are in a Hilbert space and that P_n is a self-adjoint projection (Theorem 3.4.7 in Section 3.4), we have

$$\begin{aligned} \|K(I - P_n)\| &= \left\|[K(I - P_n)]^*\right\| \\ &= \|(I - P_n)K^*\| . \end{aligned} \tag{12.3.13}$$

The first line follows from the general principle that the norm of an operator is equal to the norm of its adjoint operator. The second line follows from Theorem 3.4.7 and properties of the adjoint operation.

With Galerkin methods, it is generally the case that when P_n is regarded as an operator on the Hilbert space V, then $P_n v \to v$ for all $v \in V$. This follows if we have that the sequence of spaces $\{V_n \mid n \geq 1\}$ has the *approximating property* on V: For each $v \in V$, there is a sequence $\{v_n\}$ with $v_n \in V_n$ and

$$\lim_{n \to \infty} \|v - v_n\| = 0. \tag{12.3.14}$$

When this is combined with the optimal approximation property of Proposition 3.6.9(c), we have $P_n v \to v$ for all $v \in V$.

Recall from Lemma 2.8.13 of Chapter 2 that if K is a compact operator, then so is its adjoint K^*. Combining this with Lemma 12.1.4 and the above assumption of the pointwise convergence of P_n to I on V, we have that

$$\lim_{n \to \infty} \|(I - P_n)K^*\| = 0. \tag{12.3.15}$$

We can also apply Theorem 12.1.2 to obtain the existence and uniform boundedness of $(\lambda - P_n K)^{-1}$ for all sufficiently large n, say $n \geq N$. From (12.3.6), we also have that $(\lambda - KP_n)^{-1}$ exists and is uniformly bounded for $n \geq N$. Apply this and (12.3.12) to (12.3.11), to obtain

$$\begin{aligned} \|u - \widehat{u}_n\| &\leq \left\|(\lambda - KP_n)^{-1}\right\| \|K(I - P_n)u\| \\ &\leq c \, \|(I - P_n)K^*\| \, \|(I - P_n)u\| . \end{aligned} \tag{12.3.16}$$

Combining this with (12.3.15), we see that $\|u - \widehat{u}_n\|$ converges to zero more rapidly than does $\|(I - P_n)u\|$, or equivalently, $\|u - u_n\|$. Thus

$$\lim_{n \to \infty} \frac{\|u - \widehat{u}_n\|}{\|u - u_n\|} = 0.$$

n	$\|u-u_n\|_\infty$	*Ratio*	$\|u-\widehat{u}_n\|_\infty$	*Ratio*
2	4.66E−2		5.45E−6	
4	1.28E−2	3.6	3.48E−7	15.7
8	3.37E−3	3.8	2.19E−8	15.9

TABLE 12.2. Piecewise linear Galerkin and iterated Galerkin method for solving (12.3.18)

The quantity $\|(I-P_n)K^*\|$ can generally be estimated, in the same manner as is done for $\|(I-P_n)K\|$. Taking K to be an integral operator on $L^2(D)$, the operator K^* is an integral operator (see Example 2.6.1), with

$$K^*v(x) = \int_D k(y,x)v(y)\,dy, \quad v \in L^2(D). \tag{12.3.17}$$

Example 12.3.2 Consider the integral equation

$$\lambda u(x) - \int_0^1 e^{xy}u(y)\,dy = f(x), \quad 0 \le x \le 1 \tag{12.3.18}$$

with $\lambda = 50$ and $u(x) = e^x$. For $n \ge 1$, define the meshsize $h = 1/n$ and the mesh $x_j = jh,\ j = 0, 1, \ldots, n$. Let V_n be the set of functions which are piecewise linear on $[0,1]$ with breakpoints $x_1, \ldots, x_{n-1}$, without the continuity restriction of Section 12.2.3. The dimension of V_n is $d_n = 2n$, and this is also the order of the linear system associated with solving

$$(\lambda - P_nK)\,u_n = P_nf.$$

It is straightforward to show $\|(I-P_n)K^*\| = \mathcal{O}(h^2)$ in this case. Also, if $f \in C^2[0,1]$, then the solution u of (12.3.18) also belongs to $C^2[0,1]$; and consequently, we have $\|u - P_nu\| = \mathcal{O}(h^2)$. These results lead to

$$\|u-u_n\| = \mathcal{O}(h^2), \tag{12.3.19}$$

$$\|u-\widehat{u}_n\| = \mathcal{O}(h^4). \tag{12.3.20}$$

This is confirmed empirically in the numerical calculations which are given in Table 12.2. The error columns give the maximum error rather than the norm of the error in $L^2(0,1)$. But it can be shown that (12.3.19)–(12.3.20) generalize to $C[0,1]$ with the uniform norm. □

12.3.2 The iterated collocation solution

With collocation, the iterated solution $\widehat{u}_n$ is not always an improvement on the original collocation solution u_n, but it is for many cases of interest.

The abstract theory is still applicable, and the error equation (12.3.11) is still the focus for the error analysis:

$$u - \widehat{u}_n = (\lambda - KP_n)^{-1} K(I - P_n)u. \tag{12.3.21}$$

Recall that the projection P_n is now an interpolatory operator, as in (12.1.8). In contrast to the iterated Galerkin method, we do not have that $\|K - KP_n\|$ converges to zero. In fact, it can be shown that

$$\|K(I - P_n)\| \geq \|K\|. \tag{12.3.22}$$

To show the possibly faster convergence of $\widehat{u}_n$, we must examine collocation methods on a case-by-case basis. With some, there is an improvement. We begin with a simple example to show one of the main tools used in proving higher orders of convergence.

Consider using collocation with piecewise quadratic interpolation to solve the integral equation

$$\lambda u(x) - \int_a^b k(x,y)u(y)\,dy = f(x), \quad a \leq x \leq b. \tag{12.3.23}$$

Let $n \geq 2$ be an even integer. Define $h = (b-a)/n$ and $x_j = a + jh$, $j = 0, 1, \ldots, n$. Let V_n be the set of all continuous functions which are a quadratic polynomial when restricted to each of the subintervals $[x_0, x_2]$, $\ldots$, $[x_{n-2}, x_n]$. Easily, the dimension of V_n is $\kappa_n = n + 1$, based on each element of V_n being completely determined by its values at the $n+1$ nodes $\{x_0, \ldots, x_n\}$. Let P_n be the interpolatory projection operator from $V = C[a,b]$ to V_n.

We can write $P_n u$ in its Lagrange form:

$$P_n u(x) = \sum_{j=0}^{n} u(x_j)\ell_j(x). \tag{12.3.24}$$

For the Lagrange basis functions $\ell_j(x)$, we must distinguish the cases of even and odd indices j. For j odd,

$$\ell_j(x) = \begin{cases} -\dfrac{1}{h^2}(x - x_{j-1})(x - x_{j+1}), & x_{j-1} \leq x \leq x_{j+1}, \\ 0, & \text{otherwise.} \end{cases}$$

For j even, $2 \leq j \leq n-2$,

$$\ell_j(x) = \begin{cases} \dfrac{1}{2h^2}(x - x_{j-1})(x - x_{j-2}), & x_{j-2} \leq x \leq x_j, \\ \dfrac{1}{2h^2}(x - x_{j+1})(x - x_{j+2}), & x_j \leq x \leq x_{j+2}, \\ 0, & \text{otherwise.} \end{cases}$$

The functions $\ell_0(x)$ and $\ell_n(x)$ are appropriate modifications of this last case.

For the interpolation error on $[x_{j-2}, x_j]$, for j even, we have two formulas:

$$u(x) - P_n u(x) = (x - x_{j-2})(x - x_{j-1})(x - x_j) u[x_{j-2}, x_{j-1}, x_j, x] \quad (12.3.25)$$

and

$$u(x) - P_n u(x) = \frac{(x - x_{j-2})(x - x_{j-1})(x - x_j)}{6} u'''(c_x), \quad x_{j-2} \le x \le x_j \quad (12.3.26)$$

for some $c_x \in [x_{j-2}, x_j]$, with $u \in C^3[a, b]$. The quantity $u[x_{j-2}, x_{j-1}, x_j, x]$ is a *Newton divided difference* of order three for the function $u(x)$. From the above formulas,

$$\|u - P_n u\|_\infty \le \frac{\sqrt{3}}{27} h^3 \|u'''\|_\infty, \quad u \in C^3[a, b]. \quad (12.3.27)$$

See [15, pp. 143, 156] for details on this and more generally on divided differences.

In using piecewise quadratic functions to define the collocation method to solve (12.3.23), the result (12.3.27) implies

$$\|u - u_n\|_\infty = \mathcal{O}(h^3) \quad (12.3.28)$$

if $u \in C^3[a, b]$. To examine the error in $\widehat{u}_n$, we make a detailed examination of $K(I - P_n)u$.

Using (12.3.24),

$$K(I - P_n)u(x) = \int_a^b k(x, y) \left[u(y) - \sum_{j=0}^{n} u(x_j) \ell_j(y) \right] dy.$$

From (12.3.25),

$$K(I - P_n)u(x) = \sum_{k=1}^{n/2} \int_{x_{2k-2}}^{x_{2k}} k(x, y)(y - x_{2k-2})(y - x_{2k-1})(y - x_{2k}) \cdot u[x_{2k-2}, x_{2k-1}, x_{2k}, y] \, dy. \quad (12.3.29)$$

To examine the integral in more detail, we write it as

$$\int_{x_{2k-2}}^{x_{2k}} g_x(y) \omega(y) \, dy \quad (12.3.30)$$

with

$$\omega(y) = (y - x_{2k-2})(y - x_{2k-1})(y - x_{2k})$$

and

$$g_x(y) = k(x, y) \, u[x_{2k-2}, x_{2k-1}, x_{2k}, y].$$

Introduce

$$\nu(y) = \int_{x_{2k-2}}^{y} \omega(\xi)\, d\xi, \quad x_{2k-2} \le y \le x_{2k}.$$

Then $\nu'(y) = \omega(y)$, $\nu(y) \ge 0$ on $[x_{2k-2}, x_{2k}]$, and $\nu(x_{2k-2}) = \nu(x_{2k}) = 0$. The integral (12.3.30) becomes

$$\int_{x_{2k-2}}^{x_{2k}} g_x(y)\nu'(y)\, dy = \underbrace{\nu(y)g_x(y)|_{x_{2k-2}}^{x_{2k}}}_{=0} - \int_{x_{2k-2}}^{x_{2k}} g_x'(y)\nu(y)\, dy,$$

and so

$$\left| \int_{x_{2k-2}}^{x_{2k}} g_x'(y)\nu(y)\, dy \right| \le \|g_x'\|_\infty \int_{x_{2k-2}}^{x_{2k}} \nu(y)\, dy = \frac{4h^5}{15} \|g_x'\|_\infty .$$

In this,

$$\begin{aligned} g_x'(y) &= \frac{\partial}{\partial y} \{k(x,y)\, u[x_{2k-2}, x_{2k-1}, x_{2k}, y]\} \\ &= \frac{\partial k(x,y)}{\partial y}\, u[x_{2k-2}, x_{2k-1}, x_{2k}, y] \\ &\quad + k(x,y)\, u[x_{2k-2}, x_{2k-1}, x_{2k}, y, y]. \end{aligned}$$

The last formula uses a standard result for the differentiation of Newton divided differences (see [15, p. 147]). To have this derivation be valid, we must have $g \in C^1[a,b]$, and this is true if $u \in C^4[a,b]$ and $k_x \in C^1[a,b]$.

Combining these results, we have

$$K(I - P_n)u(x) = \mathcal{O}(h^4). \tag{12.3.31}$$

With this, we have the following theorem.

Theorem 12.3.3 *Assume that the integral equation* (12.3.23) *is uniquely solvable for all $f \in C[a,b]$. Further assume that the solution $u \in C^4[a,b]$ and that the kernel function $k(x,y)$ is continuously differentiable with respect to y. Let P_n be the interpolatory projection* (12.3.24) *defined by piecewise quadratic interpolation. Then the collocation equation $(\lambda - P_nK)u_n = P_nf$ is uniquely solvable for all sufficiently large n, say $n \ge N$; and the inverses $(\lambda - P_nK)^{-1}$ are uniformly bounded, say by $M > 0$. Moreover,*

$$\|u - u_n\|_\infty \le |\lambda|\, M \|u - P_n u\|_\infty \le \frac{\sqrt{3}\,|\lambda|\, M}{27} h^3 \|u'''\|_\infty , \quad n \ge N. \tag{12.3.32}$$

For the iterated collocation method,

$$\|u - \widehat{u}_n\|_\infty \le ch^4 \tag{12.3.33}$$

for a suitable constant $c > 0$. Consequently,

$$\max_{j=0,\dots,n} |u(x_j) - u_n(x_j)| = \mathcal{O}(h^4). \tag{12.3.34}$$

Proof. Formula (12.3.32) and the remarks preceding it are just a restatement of results from Theorem 12.1.2, applied to the particular P_n being considered here. The final bound in (12.3.32) comes from (12.3.27). The bound (12.3.33) comes from (12.3.21) and (12.3.31). The final result (12.3.34) comes from noting first that the property $P_n\widehat{u}_n = u_n$ (see (12.3.2)) implies

$$u_n(x_j) = \widehat{u}_n(x_j), \quad j = 0, \ldots, n$$

and second from applying (12.3.33). □

This theorem, and (12.3.33) in particular, shows that the iterated collocation method converges more rapidly when using the piecewise quadratic collocation method described preceding the theorem. However, when using piecewise linear interpolation to define P_n, the iterated collocation solution $\widehat{u}_n$ does not converge any more rapidly than the original solution u_n. In general, let V_n be the set of continuous piecewise polynomial functions of degree r with r an even integer, and let the collocation nodes be the breakpoints used in defining the piecewise quadratic functions. Then the iterated solution gains one extra power of h in its error bound. But this is not true if r is an odd integer.

The result (12.3.34) is an example of *superconvergence*. The rate of convergence of $u_n(x)$ at the node points $\{x_0, \ldots, x_n\}$ is greater than it is over the interval $[a, b]$ as a whole. There are a number of situations in the numerical solution of both differential and integral equations in which superconvergence occurs at special points in the domain over which the problem is defined. See Exercise 10.4.1 for a superconvergence result in the context of the finite element method. Also recall Example 12.2.1. For it, one can show that

$$K(I - P_n)\, u(x) = \mathcal{O}(h^{3/2}) \tag{12.3.35}$$

for the solution function $u(x) = \sqrt{x}$ of (12.2.9), thus proving superconvergence at the node points as was observed in Table 12.1.

The linear system for the iterated collocation solution

Let the interpolatory projection operator be written as

$$P_n u(x) = \sum_{j=1}^{\kappa_n} u(x_j)\ell_j(x), \quad u \in C(D). \tag{12.3.36}$$

When written out, the approximating equation $\lambda\widehat{u}_n - KP_n\widehat{u}_n = f$ becomes

$$\lambda\widehat{u}_n(x) - \sum_{j=1}^{\kappa_n} \widehat{u}_n(x_j) \int_D k(x, y)\ell_j(y)\, dy = f(x), \quad x \in D. \tag{12.3.37}$$

Evaluating at each node point x_i, we obtain the linear system

$$\lambda\widehat{u}_n(x_i) - \sum_{j=1}^{\kappa_n} \widehat{u}_n(x_j) \int_D k(x_i, y)\ell_j(y)\, dy = f(x_i), \quad i = 1, \ldots, \kappa_n. \tag{12.3.38}$$

This is also the linear system for the collocation solution at the node points, as given, for example, in (12.1.5) with $\phi_j = \ell_j$, or in (12.2.6). This is not surprising since u_n and $\widehat{u}_n$ agree at the node points.

The two solutions differ, however, at the remaining points in D. For general $x \in D$, $u_n(x)$ is based on the interpolation formula (12.3.36). However, the iterated collocation solution $\widehat{u}_n(x)$ is given by using (12.3.37) in the form

$$\widehat{u}_n(x) = \frac{1}{\lambda}\left[f(x) + \sum_{j=1}^{\kappa_n} \widehat{u}_n(x_j) \int_D k(x,y)\ell_j(y)\,dy\right], \quad x \in D. \tag{12.3.39}$$

In Section 12.5, we see that (12.3.37)–(12.3.39) is a special case of the Nyström method for solving $(\lambda - K)\,u = f$, and (12.3.39) is called the Nyström interpolation function. Generally, it is more accurate than ordinary polynomial interpolation.

Exercise 12.3.1 Prove (12.3.10).

Exercise 12.3.2 Consider the piecewise linear Galerkin scheme of Subsection 12.2.3. Assume that $k(y,x)$ is twice continuously differentiable with respect to y, for $a \le y, x \le b$. Analyze the convergence of the iterated Galerkin method for this case. What is the rate of convergence of the iterated Galerkin solutions $\{\widehat{u}_n\}$?

Exercise 12.3.3 Derive the identity

$$\lambda(u - \widehat{u}_n) = K(I - P_n)u + KP_n(u - \widehat{u}_n)$$

for the solution u of $(\lambda - K)\,u = f$ and the iterated projection solutions $\{\widehat{u}_n\}$. Using this, obtain results on the uniform convergence of $\{\widehat{u}_n\}$ and $\{u_n\}$ to u for the integral operator K of (12.2.1).
Hint: Write $K(I - P_n)u$ and $KP_n(u - \widehat{u}_n)$ as an inner products to which the Cauchy-Schwarz inequality can be applied.

Exercise 12.3.4 Prove (12.3.22).
Hint: Look at functions v for which $\|v\| = 1$ and $\|K\| \approx \|Kv\|$. Then modify v to w with $P_n w$ and $\|Kv\| \approx \|Kw\|$.

Exercise 12.3.5 Let $n > 0$, $h = (b-a)/n$, and $\tau_j = a + jh$ for $j = 0, 1, \ldots, n$. Let V_n be the set of piecewise linear functions, linear over each subinterval $[\tau_{j-1}, \tau_j]$, with no restriction that the functions be continuous. The dimension of V_n is $2n$. To define the collocation nodes, introduce

$$\mu_1 = \frac{3 - \sqrt{3}}{6}, \quad \mu_2 = \frac{3 + \sqrt{3}}{6}.$$

On each subinterval $[\tau_{j-1}, \tau_j]$, define two collocation nodes by

$$x_{2j-1} = \tau_{j-1} + h\mu_1, \quad x_{2j} = \tau_{j-1} + h\mu_2.$$

Define a collocation method for solving (12.3.23) using the approximating subspace V_n and the collocation nodes $\{x_1, \ldots, x_{2n}\}$. Assume $u \in C^4[a,b]$ and assume $k(x,y)$ is twice continuously differentiable with respect to y, for $a \le y, x \le b$. It can be shown from the methods of Section 12.5 that $(\lambda - P_n K)^{-1}$ exists and is uniformly bounded for all sufficiently large n, say $n \ge N$. Assuming this, show that

$$\begin{aligned} \|u - u_n\|_\infty &= \mathcal{O}(h^2), \\ \|u - \widehat{u}_n\|_\infty &= \mathcal{O}(h^4), \\ \max_{j=1,\ldots,2n} |u(x_j) - u_n(x_j)| &= \mathcal{O}(h^4). \end{aligned}$$

Hint: The polynomial $(\mu - \mu_1)(\mu - \mu_2)$ is the Legendre polynomial of degree 2 on $[0,1]$, and therefore it is orthogonal over $[0,1]$ to all polynomials of lesser degree.

Exercise 12.3.6 Consider the integral equation $(\lambda - K)u = f$, with

$$Kv(x) = \int_a^b k(x,y)v(y)\,dy, \quad y \in [a,b],\ v \in C[a,b].$$

Assume that u is continuous, but that its first derivative is discontinuous. Moreover, assume the kernel function $k(x,y)$ is several times continuously differentiable with respect to x. Write

$$u = \frac{1}{\lambda}(f + Ku) \equiv \frac{1}{\lambda}(f + w).$$

Prove that w satisfies the equation $(\lambda - K)w = Kf$. Show that w is smoother than u. Be as precise as possible in stating your results.

Exercise 12.3.7 (continuation of Exercise 12.3.6). Apply a projection method to the solution of the modified equation $(\lambda - K)w = Kf$, denoting the approximate solution by w_n. Then define $u_n = (f + w_n)/\lambda$. Analyze the convergence of $\{u_n\}$. Compare the method to the original projection method applied directly to $(\lambda - K)u = f$.

Exercise 12.3.8 Derive (12.3.35) in the case the integral operator K has the kernel function $k \equiv 1$.

Exercise 12.3.9 Generalize Exercise 12.3.8 to the case of a general kernel function $k(x,y)$ that is once continuously differentiable with respect to the variable y on the interval of integration.

12.4 The Nyström method

The Nyström method was originally introduced to handle approximations based on numerical integration of the integral operator in the equation

$$\lambda u(x) - \int_D k(x,y)u(y)\,dy = f(x), \quad x \in D. \tag{12.4.1}$$

The resulting solution is found first at the set of quadrature node points, and then it is extended to all points in D by means of a special, and generally quite accurate, interpolation formula. The numerical method is much simpler to implement on a computer, but the error analysis is more sophisticated than for projection methods. The resulting theory has taken an abstract form which also includes an error analysis of projection methods, although the latter are probably still best understood as distinct methods of interest in their own right.

12.4.1 The Nyström method for continuous kernel functions

Let a numerical integration scheme be given:

$$\int_D g(y)\,dy \approx \sum_{j=1}^{q_n} w_{n,j} g(x_{n,j}), \quad g \in C(D) \tag{12.4.2}$$

with an increasing sequence of values of n. We assume that for every $g \in C(D)$, the numerical integrals converge to the true integral as $n \to \infty$. As in Subsection 2.4.4, this implies

$$c_I \equiv \sup_{n \geq 1} \sum_{j=1}^{q_n} |w_{n,j}| < \infty. \tag{12.4.3}$$

To simplify the notation, we omit the subscript n, so that $w_{n,j} \equiv w_j$, $x_{n,j} \equiv x_j$; but the presence of n is to be understood implicitly. On occasion, we also use $q \equiv q_n$.

Let $k(x, y)$ be continuous for all $x, y \in D$, where D is a closed and bounded set in $\mathbb{R}^d$ for some $d \geq 1$. Usually, in fact, we want $k(x, y)$ to be several times continuously differentiable. Using the above quadrature scheme, approximate the integral in (12.4.1), obtaining a new equation:

$$\lambda u_n(x) - \sum_{j=1}^{q_n} w_j k(x, x_j) u_n(x_j) = f(x), \quad x \in D. \tag{12.4.4}$$

We write this as an exact equation with a new unknown function $u_n(x)$. To find the solution at the node points, let x run through the quadrature node points x_i. This yields

$$\lambda u_n(x_i) - \sum_{j=1}^{q_n} w_j k(x_i, x_j) u_n(x_j) = f(x_i), \quad i = 1, \ldots, q_n \tag{12.4.5}$$

which is a linear system of order q_n. The unknown is a vector

$$\underline{u}_n \equiv (u_n(x_1), \ldots, u_n(x_q))^T.$$

Each solution $u_n(x)$ of (12.4.4) furnishes a solution to (12.4.5): merely evaluate $u_n(x)$ at the node points. The converse is also true. To each solution $\underline{z} \equiv [z_1, \ldots, z_q]^T$ of (12.4.5), there is a unique solution of (12.4.4) which agrees with $\underline{z}$ at the node points. If one solves for $u_n(x)$ in (12.4.4), then $u_n(x)$ is determined by its values at the node points $\{x_j\}$. Therefore, when given a solution $\underline{z}$ to (12.4.5), define

$$z(x) = \frac{1}{\lambda}\left[f(x) + \sum_{j=1}^{q_n} w_j k(x, x_j) z_j\right], \quad x \in D. \tag{12.4.6}$$

This is an interpolation formula. In fact,

$$z(x_i) = \frac{1}{\lambda}\left[f(x_i) + \sum_{j=1}^{q_n} w_j k(x_i, x_j) z_j\right] = z_i$$

for $i = 1, \ldots, q_n$. The last step follows from $\underline{z}$ being a solution to (12.4.5). Using this interpolation result in (12.4.6), we have that $z(x)$ solves (12.4.4). The uniqueness of the relationship between $\underline{z}$ and $z(x)$ follows from the solutions $u_n(x)$ of (12.4.4) being completely determined by their values at the nodes $\{x_i\}$.

The formula (12.4.6) is called the *Nyström interpolation formula*. In the original paper of Nyström [182], the author uses a highly accurate Gaussian quadrature formula with a very small number of quadrature nodes (e.g. $q = 3$). He then uses (12.4.6) in order to extend the solution to all other $x \in D$ while retaining the accuracy found in the solution at the node points. The formula (12.4.6) is usually a very good interpolation formula.

Example 12.4.1 Consider the integral equation

$$\lambda u(x) - \int_0^1 e^{yx} u(y)\, dy = f(x), \quad 0 \le x \le 1, \tag{12.4.7}$$

with $\lambda = 2$ and $u(x) = e^x$. Since $\|K\| = e - 1 \doteq 1.72$, the geometric series theorem (Theorem 2.3.1 in Chapter 2) implies the integral equation is uniquely solvable for any given $f \in C[0, 1]$.

Consider first using the three-point Simpson rule to approximate (12.4.7), with nodes $\{0, 0.5, 1\}$. Then the errors at the nodes are collectively

$$\begin{pmatrix} u(0) \\ u(.5) \\ u(1) \end{pmatrix} - \begin{pmatrix} u_3(0) \\ u_3(.5) \\ u_3(1) \end{pmatrix} \doteq \begin{pmatrix} -0.0047 \\ -0.0080 \\ -0.0164 \end{pmatrix} \tag{12.4.8}$$

which are reasonably small errors. For comparison, use Gauss-Legendre quadrature with 3 nodes,

$$\int_0^1 g(x)\, dx \approx \frac{1}{18}\left[5g(x_1) + 8g(x_2) + 5g(x_3)\right]$$

where

$$x_1 = \frac{1-\sqrt{0.6}}{2} \doteq 0.11270167, \quad x_2 = 0.5, \quad x_3 = \frac{1+\sqrt{0.6}}{2} \doteq 0.88729833.$$

The errors at the nodes in solving (12.4.7) with the Nyström method are now collectively

$$\begin{pmatrix} u(x_1) \\ u(x_2) \\ u(x_3) \end{pmatrix} - \begin{pmatrix} u_3(x_1) \\ u_3(x_2) \\ u_3(x_3) \end{pmatrix} \doteq \begin{pmatrix} 2.10 \times 10^{-5} \\ 3.20 \times 10^{-5} \\ 6.32 \times 10^{-5} \end{pmatrix} \tag{12.4.9}$$

and these errors are much smaller than with Simpson's rule when using an equal number of node points. Generally, Gaussian quadrature is much superior to Simpson's rule; but it results in the answers being given at the Gauss-Legendre nodes, which is usually not a convenient choice for subsequent uses of the answers.

Quadratic interpolation can be used to extend the numerical solution to all other $x \in [0, 1]$, but it generally results in much larger errors. For example,

$$u(1.0) - P_2 u_3(1.0) \doteq 0.0158$$

where $P_2 u_3(x)$ denotes the quadratic polynomial interpolating the Nyström solution at the Gaussian quadrature node points given above. In contrast, the Nyström formula (12.4.6) gives errors which are consistent in size with those in (12.4.9). For example,

$$u(1.0) - u_3(1.0) \doteq 8.08 \times 10^{-5}.$$

A graph of the error in $u_3(x)$ over $[0, 1]$ is given in Figure 12.1, with the errors at the node points indicated by '$\diamond$'. □

12.4.2 Properties and error analysis of the Nyström method

The Nyström method is implemented with the finite linear system (12.4.5), but the formal error analysis is done using the functional equation (12.4.4). As before, we write the integral equation (12.4.1) in abstract form as

$$(\lambda - K)\, u = f;$$

and we write the numerical integral equation (12.4.4) as

$$(\lambda - K_n) u_n = f.$$

The Banach space for our initial error analysis is $V = C(D)$. The numerical integral operator

$$K_n u(x) \equiv \sum_{j=1}^{q_n} w_j k(x, x_j) u(x_j), \quad x \in D,\ u \in C(D), \tag{12.4.10}$$

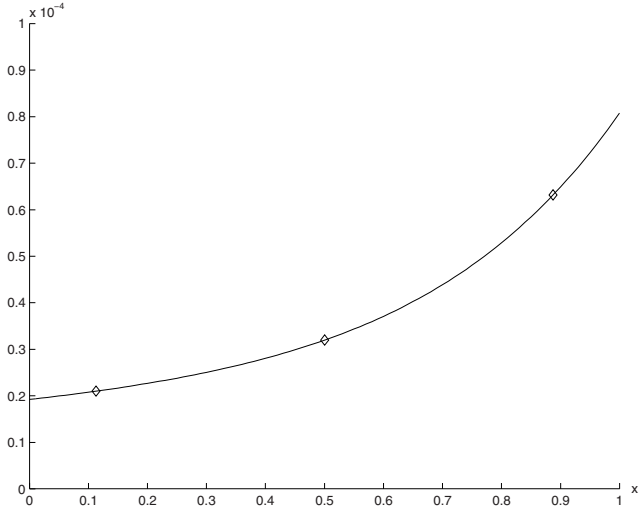

FIGURE 12.1. Error in Nyström interpolation with 3 point Gauss-Legendre quadrature

is a bounded, finite rank linear operator on $C(D)$ to $C(D)$, with

$$\|K_n\| = \max_{x\in D} \sum_{j=1}^{q_n} |w_j k(x, x_j)| \,. \tag{12.4.11}$$

The error analyses of projection methods depended on showing $\|K-K_n\|$ converges to zero as n increases, with $K_n = P_n K$ the approximation to the integral operator K. This cannot be done here; and in fact,

$$\|K - K_n\| \geq \|K\|. \tag{12.4.12}$$

We leave the proof of this as an exercise for the reader. Because of this result, the standard type of perturbation analysis which was used earlier needs to be modified. We begin by looking at quantities which do converge to zero as $n \to \infty$.

Lemma 12.4.2 *Let D be a closed, bounded set in $\mathbb{R}^d$; and let $k(x, y)$ be continuous for $x, y \in D$. Let the quadrature scheme* (12.4.2) *be convergent for all continuous functions on D. Define*

$$e_n(x, y) = \int_D k(x, v)k(v, y)\, dv - \sum_{j=1}^{q_n} w_j k(x, x_j)k(x_j, y)\,, \;\; x, y \in D, \;\; n \geq 1, \tag{12.4.13}$$

the numerical integration error for the integrand $k(x,\cdot)k(\cdot,y)$. Then for $z \in C(D)$,

$$(K - K_n)Kz(x) = \int_D e_n(x,y)z(y)\,dy, \tag{12.4.14}$$

$$(K - K_n)K_n z(x) = \sum_{j=1}^{q_n} w_j e_n(x,x_j)z(x_j). \tag{12.4.15}$$

In addition,

$$\|(K - K_n)K\| = \max_{x\in D} \int_D |e_n(x,y)|\,dy, \tag{12.4.16}$$

$$\|(K - K_n)K_n\| = \max_{x\in D} \sum_{j=1}^{q_n} |w_j e_n(x,x_j)|\,. \tag{12.4.17}$$

Finally, the numerical integration error e_n converges to zero uniformly on D,

$$c_E \equiv \lim_{n\to\infty} \max_{x,y\in D} |e_n(x,y)| = 0 \tag{12.4.18}$$

and thus

$$\|(K - K_n)K\|,\ \|(K - K_n)K_n\| \to 0 \quad \textit{as } n \to \infty. \tag{12.4.19}$$

Proof. The proofs of (12.4.14) and (12.4.15) are straightforward manipulations and we omit them. The quantity $(K - K_n)K$ is an integral operator on $C(D)$, by (12.4.14); and therefore, we have (12.4.16) for its bound. The proof of (12.4.17) is also straightforward and we omit it.

To prove (12.4.18), we begin by showing that $\{e_n(x,y) \mid n \geq 1\}$ is a uniformly bounded and equicontinuous family which is pointwise convergent to 0 on the closed bounded set D; and then $e_n(x,y) \to 0$ uniformly on D by the Ascoli Theorem. By the assumption that the quadrature rule of (12.4.2) converges for all continuous functions g on D, we have that for each $x, y \in D,\ e_n(x,y) \to 0$ as $n \to \infty$.

To prove boundedness,

$$|e_n(x,y)| \leq (c_D + c_I)\,c_K^2$$

with

$$c_D = \int_D dy, \quad c_K = \max_{x,y\in D} |k(x,y)|$$

and c_I the constant from (12.4.3). For equicontinuity,

$$\begin{aligned}
|e_n(x,y) - e_n(\xi,\eta)| &\leq |e_n(x,y) - e_n(\xi,y)| + |e_n(\xi,y) - e_n(\xi,\eta)|\,,\\
|e_n(x,y) - e_n(\xi,y)| &\leq c_K(c_D + c_I) \max_{y\in D} |k(x,y) - k(\xi,y)|\,,\\
|e_n(\xi,y) - e_n(\xi,\eta)| &\leq c_K(c_D + c_I) \max_{x\in D} |k(x,y) - k(x,\eta)|\,.
\end{aligned}$$

By the uniform continuity of $k(x,y)$ on the closed bounded set D, this shows the equicontinuity of $\{e_n(x,y)\}$. This also completes the proof of (12.4.18).

For (12.4.19) we notice that

$$\|(K-K_n)K\| \leq c_D \max_{x,y\in D} |e_n(x,y)|\,, \tag{12.4.20}$$

$$\|(K-K_n)K_n\| \leq c_I \max_{x,y\in D} |e_n(x,y)|\,. \tag{12.4.21}$$

This completes the proof. □

To carry out an error analysis for the Nyström method (12.4.4)–(12.4.6), we need the following perturbation theorem. It furnishes an alternative to the perturbation arguments based on the geometric series theorem (e.g. Theorem 2.3.5 in Section 2.3).

Theorem 12.4.3 *Let V be a Banach space, let S,T be bounded operators on V to V, and let S be compact. For given $\lambda \neq 0$, assume $\lambda - T : V \xrightarrow[onto]{1-1} V$, which implies $(\lambda - T)^{-1}$ exists as a bounded operator on V to V. Finally, assume*

$$\|(T-S)S\| < \frac{|\lambda|}{\|(\lambda-T)^{-1}\|}. \tag{12.4.22}$$

Then $(\lambda - S)^{-1}$ exists and is bounded on V to V, with

$$\left\|(\lambda-S)^{-1}\right\| \leq \frac{1+\left\|(\lambda-T)^{-1}\right\|\|S\|}{|\lambda| - \|(\lambda-T)^{-1}\|\,\|(T-S)S\|}. \tag{12.4.23}$$

If $(\lambda - T)u = f$ and $(\lambda - S)z = f$, then

$$\|u-z\| \leq \left\|(\lambda-S)^{-1}\right\|\|Tu-Su\|\,. \tag{12.4.24}$$

Proof. Consider that if $(\lambda - S)^{-1}$ were to exist, then it would satisfy the identity

$$(\lambda-S)^{-1} = \frac{1}{\lambda}\left[I + (\lambda-S)^{-1}S\right]. \tag{12.4.25}$$

Without any motivation at this point, consider the approximation

$$(\lambda-S)^{-1} \approx \frac{1}{\lambda}\left[I + (\lambda-T)^{-1}S\right]. \tag{12.4.26}$$

To check this approximation, compute

$$\frac{1}{\lambda}\left[I + (\lambda-T)^{-1}S\right](\lambda-S) = I + \frac{1}{\lambda}(\lambda-T)^{-1}(T-S)S. \tag{12.4.27}$$

The right side is invertible by the geometric series theorem, because (12.4.22) implies

$$\frac{1}{|\lambda|}\left\|(\lambda-T)^{-1}\right\|\|(T-S)S\| < 1.$$

In addition, the geometric series theorem implies, after simplification, that

$$\left\|\left[\lambda+(\lambda-T)^{-1}(T-S)S\right]^{-1}\right\| \leq \frac{1}{|\lambda|-\|(\lambda-T)^{-1}\|\,\|(T-S)S\|}. \tag{12.4.28}$$

Since the right side of (12.4.27) is invertible, the left side is also invertible. This implies that $\lambda - S$ is one-to-one, as otherwise the left side would not be invertible. Since S is compact, the Fredholm Alternative Theorem (Theorem 2.8.10 of Subsection 2.8.4) implies $(\lambda - S)^{-1}$ exists and is bounded on V to V. In particular,

$$(\lambda-S)^{-1} = \left[\lambda+(\lambda-T)^{-1}(T-S)S\right]^{-1}\left[I+(\lambda-T)^{-1}S\right]. \tag{12.4.29}$$

The bound (12.4.23) follows directly from this and (12.4.28).

For the error $u - z$, rewrite $(\lambda - T)u = f$ as

$$(\lambda-S)u = f + (T-S)u.$$

Subtract $(\lambda - S)z = f$ to get

$$(\lambda-S)(u-z) = (T-S)u, \tag{12.4.30}$$

$$u-z = (\lambda-S)^{-1}(T-S)u. \tag{12.4.31}$$

Take the norm,

$$\|u-z\| \leq \left\|(\lambda-S)^{-1}\right\|\,\|(T-S)u\|,$$

which proves (12.4.24). □

Using this theorem, we can give a complete convergence analysis for the Nyström method (12.4.4)–(12.4.6).

Theorem 12.4.4 *Let D be a closed, bounded set in $\mathbb{R}^d$; and let $k(x, y)$ be continuous for $x, y \in D$. Assume the quadrature scheme* (12.4.2) *is convergent for all continuous functions on D. Further, assume that the integral equation* (12.4.1) *is uniquely solvable for given $f \in C(D)$, with $\lambda \neq 0$. Then for all sufficiently large n, say $n \geq N$, the approximate inverses $(\lambda - K_n)^{-1}$ exist and are uniformly bounded,*

$$\left\|(\lambda-K_n)^{-1}\right\| \leq \frac{1+\left\|(\lambda-K)^{-1}\right\|\,\|K_n\|}{|\lambda|-\|(\lambda-K)^{-1}\|\,\|(K-K_n)K_n\|} \leq c_y, \quad n \geq N, \tag{12.4.32}$$

with a suitable constant $c_y < \infty$. For the equations $(\lambda - K)u = f$ and $(\lambda - K_n)u_n = f$, we have

$$\begin{aligned}\|u-u_n\|_\infty &\leq \left\|(\lambda-K_n)^{-1}\right\|\,\|(K-K_n)u\|_\infty \\ &\leq c_y\,\|(K-K_n)u\|_\infty, \qquad n \geq N.\end{aligned} \tag{12.4.33}$$

Proof. The proof is a simple application of the preceding theorem, with $S = K_n$ and $T = K$. From Lemma 12.4.2, we have $\|(K - K_n)K_n\| \to 0$, and therefore (12.4.22) is satisfied for all sufficiently large n, say $n \geq N$. From (12.4.11), the boundedness of $k(x, y)$ over D, and (12.4.3),

$$\|K_n\| \leq c_I c_K, \quad n \geq 1.$$

Then

$$c_y \equiv \sup_{n \geq N} \frac{1 + \left\|(\lambda - K)^{-1}\right\| \|K_n\|}{|\lambda| - \|(\lambda - K)^{-1}\| \|(K - K_n)K_n\|} < \infty. \tag{12.4.34}$$

This completes the proof. □

This last theorem gives complete information for analyzing the convergence of the Nyström method (12.4.4)–(12.4.6). The term $\|(K - K_n)K_n\|$ can be analyzed from (12.4.17) by analyzing the numerical integration error $e_n(x, y)$ of (12.4.13). From the error bound (12.4.33), the speed with which $\|u - u_n\|_\infty$ converges to zero is bounded by that of the numerical integration error

$$\|(K - K_n)u\|_\infty = \max_{x \in D} \left| \int_D k(x, y)u(y)\, dy - \sum_{j=1}^{q_n} w_j k(x, x_j)u(x_j) \right|. \tag{12.4.35}$$

In fact, the error $\|u - u_n\|_\infty$ converges to zero with exactly this speed. Recall from applying (12.4.30) that

$$(\lambda - K_n)(u - u_n) = (K - K_n)u. \tag{12.4.36}$$

From bounding this,

$$\|(K - K_n)u\|_\infty \leq \|\lambda - K_n\| \|u - u_n\|_\infty.$$

When combined with (12.4.33), this shows the assertion that $\|u - u_n\|_\infty$ and $\|(K - K_n)u\|_\infty$ converge to zero with the same speed.

There is a very large literature on bounding and estimating the errors for the common numerical integration rules. Thus the speed of convergence with which $\|u - u_n\|_\infty$ converges to zero can be determined by using results on the speed of convergence of the integration rule (12.4.2) when it is applied to the integral

$$\int_D k(x, y)u(y)\, dy.$$

Example 12.4.5 Consider the trapezoidal numerical integration rule

$$\int_a^b g(y)\, dy \approx h \sum_{j=0}^{n}{}'' g(x_j) \tag{12.4.37}$$

with $h = (b-a)/n$ and $x_j = a + jh$ for $j = 0, \ldots, n$. The notation Σ'' means the first and last terms are to be halved before summing. For the error,

$$\int_a^b g(y)\,dy - h\sum_{j=0}^{n}{}''g(x_j) = -\frac{h^2(b-a)}{12}g''(\xi_n), \quad g \in C^2[a,b],\ n \geq 1 \tag{12.4.38}$$

with ξ_n some point in $[a,b]$. There is also the asymptotic error formula

$$\int_a^b g(y)\,dy - h\sum_{j=0}^{n}{}''g(x_j) = -\frac{h^2}{12}\left[g'(b) - g'(a)\right] + \mathcal{O}(h^4), \quad g \in C^4[a,b] \tag{12.4.39}$$

and we make use of it in a later example. For a derivation of these formulas, see [15, p. 285].

When this is applied to the integral equation

$$\lambda u(x) - \int_a^b k(x,y)u(y)\,dy = f(x), \quad a \leq x \leq b, \tag{12.4.40}$$

we obtain the approximating linear system

$$\lambda u_n(x_i) - h\sum_{j=0}^{n}{}''k(x_i,x_j)u_n(x_j) = f(x_i), \quad i = 0,1,\ldots,n, \tag{12.4.41}$$

which is of order $q_n = n+1$. The Nyström interpolation formula is given by

$$u_n(x) = \frac{1}{\lambda}\left[f(x) + h\sum_{j=0}^{n}{}''k(x,x_j)u_n(x_j)\right], \quad a \leq x \leq b. \tag{12.4.42}$$

The speed of convergence is based on the numerical integration error

$$(K - K_n)u(y) = -\frac{h^2(b-a)}{12}\left.\frac{\partial^2 k(x,y)u(y)}{\partial y^2}\right|_{y=\xi_n(x)} \tag{12.4.43}$$

with $\xi_n(x) \in [a,b]$. From (12.4.39), the asymptotic integration error is

$$(K - K_n)u(y) = -\frac{h^2}{12}\left.\frac{\partial k(x,y)u(y)}{\partial y}\right|_{y=a}^{y=b} + \mathcal{O}(h^4). \tag{12.4.44}$$

From (12.4.43), we see the Nyström method converges with an order of $\mathcal{O}(h^2)$, provided $k(x,y)u(y)$ is twice continuously differentiable with respect to y, uniformly in x. □

An asymptotic error estimate

In those cases for which the quadrature formula has an asymptotic error formula, as in (12.4.39), we can give an asymptotic estimate of the error in solving the integral equation using the Nyström method. Returning to (12.4.36), we can write

$$u - u_n = (\lambda - K_n)^{-1}(K - K_n)u = \epsilon_n + r_n \tag{12.4.45}$$

with

$$\epsilon_n = (\lambda - K)^{-1}(K - K_n)u$$

and

$$\begin{aligned} r_n &= \left[(\lambda - K_n)^{-1} - (\lambda - K)^{-1}\right](K - K_n)\, u \\ &= (\lambda - K_n)^{-1}(K_n - K)(\lambda - K)^{-1}(K - K_n)\, u. \end{aligned} \tag{12.4.46}$$

The term r_n generally converges to zero more rapidly than the term ϵ_n, although showing this is dependent on the quadrature rule being used. Assuming the latter to be true, we have

$$u - u_n \approx \epsilon_n \tag{12.4.47}$$

with ϵ_n satisfying the original integral equation with the integration error $(K - K_n)u$ as the right hand side,

$$(\lambda - K)\epsilon_n = (K - K_n)u. \tag{12.4.48}$$

At this point, one needs to consider the quadrature rule in more detail.

Example 12.4.6 Consider again the earlier example (12.4.37)–(12.4.44) of the Nyström method with the trapezoidal rule. Assume further that $k(x, y)$ is four times continuously differentiable with respect to both y and x, and assume $u \in C^4[a, b]$. Then from the asymptotic error formula (12.4.44), we can decompose the right side $(K - K_n)\, u$ of (12.4.48) into two terms, of sizes $\mathcal{O}(h^2)$ and $\mathcal{O}(h^4)$. Introduce the function $\gamma(y)$ satisfying the integral equation

$$\lambda\gamma(x) - \int_a^b k(x, y)\gamma(y)\, dy = -\frac{1}{12}\, \frac{\partial k(x, y)u(y)}{\partial y}\bigg|_{y=a}^{y=b}, \quad a \le x \le b.$$

Then the error term ϵ_n in (12.4.47)–(12.4.48) is dominated by $\gamma(x)h^2$. By a similar argument, it can also be shown that the term $r_n = \mathcal{O}(h^4)$. Thus we have the asymptotic error estimate

$$u - u_n \approx \gamma(x)h^2 \tag{12.4.49}$$

for the Nyström method with the trapezoidal rule. □

Conditioning of the linear system

Let A_n denote the matrix of coefficients for the linear system (12.4.5):

$$(A_n)_{i,j} = \lambda\delta_{i,j} - w_j k(x_i, x_j).$$

We want to bound $\text{cond}(A_n) = \|A_n\| \, \|A_n^{-1}\|$.

For general $z \in C(D)$,

$$\begin{aligned}\max_{i=1,\dots,q_n} \left|\lambda z(x_i) - \sum_{j=1}^{q_n} w_j k(x_i, x_j) z(x_j)\right| \\ \leq \sup_{x\in D} \left|\lambda z(x) - \sum_{j=1}^{q_n} w_j k(x, x_j) z(x_j)\right|.\end{aligned}$$

This shows

$$\|A_n\| \leq \|\lambda - K_n\|. \tag{12.4.50}$$

For A_n^{-1},

$$\left\|A_n^{-1}\right\| = \sup_{\substack{\gamma\in\mathbb{R}^{q_n} \\ \|\gamma\|_\infty = 1}} \left\|A_n^{-1}\gamma\right\|_\infty .$$

For such γ, let $z = A_n^{-1}\gamma$ or $\gamma = A_n z$. Pick $f \in C(D)$ such that

$$f(x_i) = \gamma_i, \quad i = 1, \dots, q_n$$

and $\|f\|_\infty = \|\gamma\|_\infty$. Let $u_n = (\lambda - K_n)^{-1} f$, or equivalently, $(\lambda - K_n)\, u_n = f$. Then from the earlier discussion of the Nyström method,

$$u_n(x_i) = z_i, \quad i = 1, \dots, q_n.$$

Then

$$\begin{aligned}\left\|A_n^{-1}\gamma\right\|_\infty &= \|z\|_\infty \\ &\leq \|u_n\|_\infty \\ &\leq \left\|(\lambda - K_n)^{-1}\right\| \|f\|_\infty \\ &= \left\|(\lambda - K_n)^{-1}\right\| \|\gamma\|_\infty .\end{aligned}$$

This proves

$$\left\|A_n^{-1}\right\|_\infty \leq \left\|(\lambda - K_n)^{-1}\right\|. \tag{12.4.51}$$

Combining these results,

$$\text{cond}(A_n) \leq \|\lambda - K_n\| \left\|(\lambda - K_n)^{-1}\right\| \equiv \text{cond}(\lambda - K_n). \tag{12.4.52}$$

Thus if the operator equation $(\lambda - K_n)u_n = f$ is well-conditioned, then so is the linear system associated with it. We leave as an exercise the development of the relationship between $\text{cond}(\lambda - K_n)$ and $\text{cond}(\lambda - K)$.

12.4.3 Collectively compact operator approximations

The error analysis of the Nyström method was developed mainly during the period 1940 to 1970, and a number of researchers were involved. Initially, the only goal was to show that the method was stable and convergent, and perhaps, to obtain computable error bounds. As this was accomplished, a second goal emerged of creating an abstract framework for the method and its error analysis, a framework in the language of functional analysis which referred only to mapping properties of the approximate operators and not to properties of the particular integral operator, function space, or quadrature scheme being used. The final framework developed is due primarily to P. Anselone, and he gave to it the name of the *theory of collectively compact operator approximations*. A complete presentation of it is given in his book [5], and we present only a portion of it here. With this framework, it has been possible to analyze a number of important extensions of the Nyström method, including those discussed in the following Section 12.5. For an extended discussion of this theory, see [19].

Within a functional analysis framework, how does one characterize the numerical integral operators $\{K_n \mid n \geq 1\}$? We want to know the characteristic properties of these operators which imply that $\|(K - K_n)K_n\| \to 0$ as $n \to \infty$. Then the earlier Theorem 12.4.3 remains valid, and the Nyström method and its error analysis can be extended to other situations, some of which are discussed in later sections.

We assume that $\{K_n \mid n \geq 1\}$ satisfies the following properties.

A1. V is a Banach space; and K and K_n, $n \geq 1$, are linear operators on V into V.

A2. $K_n u \to Ku$ as $n \to \infty$, for all $u \in V$.

A3. The set $\{K_n \mid n \geq 1\}$ is *collectively compact*, which means that the set

$$S = \{K_n v \mid n \geq 1 \text{ and } \|v\| \leq 1\} \tag{12.4.53}$$

has compact closure in V.

These assumptions are an excellent abstract characterization of the numerical integral operators introduced earlier in (12.4.10) of this chapter. We refer to a family $\{K_n\}$ which satisfies **A1**–**A3** as a *collectively compact family of pointwise convergent operators*.

Lemma 12.4.7 *Assume the above properties* **A1**–**A3**. *Then:*

1. *K is compact;*
2. *$\{K_n \mid n \geq 1\}$ is uniformly bounded;*

3. For any compact operator $M : V \to V$,

$$\|(K - K_n)M\| \to 0 \quad \textit{as } n \to \infty;$$

4. $\|(K - K_n)K_n\| \to 0$ *as* $n \to \infty$.

Proof. (1) To show K is compact, it is sufficient to show that the set

$$\{Kv \mid \|v\| \leq 1\}$$

has compact closure in V. By **A2**, this last set is contained in $\overline{S}$, and it is compact by **A3**.
(2) This follows from the definition of operator norm and the boundedness of the set $\overline{S}$.
(3) Using the definition of operator norm,

$$\begin{aligned} \|(K - K_n)M\| &= \sup_{\|v\| \leq 1} \|(K - K_n)Mu\| \\ &= \sup_{z \in M(B)} \|(K - K_n)z\| \end{aligned} \tag{12.4.54}$$

with $B = \{v \mid \|v\| \leq 1\}$. From the compactness of M, the set $M(B)$ has compact closure. Using Lemma 12.4.2, we then have that the last quantity in (12.4.54) goes to zero as $n \to \infty$.
(4) Again, using the definition of operator norm,

$$\|(K - K_n)K_n\| = \sup_{\|v\| \leq 1} \|(K - K_n)K_n v\| = \sup_{z \in S} \|(K - K_n)z\| . \tag{12.4.55}$$

Using **A3**, S has compact closure; and then using Lemma 12.4.2, we have that the last quantity in (12.4.55) goes to zero as $n \to \infty$. □

As a consequence of this lemma, we can apply Theorem 12.4.3 to any set of approximating equations $(\lambda - K_n)u_n = f$ where the set $\{K_n\}$ satisfies **A1**–**A3**. This extends the idea of the Nyström method, and the product integration methods of the following section is analyzed using this more abstract framework.

Returning to the proof of Theorem 12.4.3, we can better motivate an argument used there. With $S = K_n$ and $T = K$, the statements (12.4.25) and (12.4.26) become

$$(\lambda - K_n)^{-1} = \frac{1}{\lambda}\left[I + (\lambda - K_n)^{-1}K_n\right], \tag{12.4.56}$$

$$(\lambda - K_n)^{-1} \approx \frac{1}{\lambda}\left[I + (\lambda - K)^{-1}K_n\right]. \tag{12.4.57}$$

Since K_n is not norm convergent to K, we cannot expect $(\lambda - K)^{-1} \approx (\lambda - K_n)^{-1}$ to be a good approximation. However, it becomes a much better approximation when the operators are restricted to act on a compact subset of V. Since the family $\{K_n\}$ is collectively compact, (12.4.57) is a good approximation of (12.4.56).

Exercise 12.4.1 Prove (12.4.12).
Hint: Recall the discussion in Exercise 12.3.4.

Exercise 12.4.2 Derive (12.4.14)–(12.4.17).

Exercise 12.4.3 Obtain a bound for $\text{cond}(\lambda - K_n)$ in terms of $\text{cond}(\lambda - K)$. More generally, explore the relationship between these two condition numbers.
Hint: Use Theorem 12.4.4.

Exercise 12.4.4 Generalize Example 12.4.5 to Simpson's rule.

Exercise 12.4.5 Generalize Example 12.4.6 to Simpson's rule.

12.5 Product integration

We now consider the numerical solution of integral equations of the second kind in which the kernel function $k(x, y)$ is not continuous, but for which the associated integral operator K is still compact on $C(D)$ into $C(D)$. The main ideas we present will extend to functions in any finite number of variables; but it is more intuitive to first present these ideas for integral equations for functions of a single variable, and in particular,

$$\lambda u(x) - \int_a^b k(x, y)u(y)\, dy = f(x), \quad a \le x \le b. \tag{12.5.1}$$

In this setting, most such discontinuous kernel functions $k(x, y)$ have an infinite singularity; and the most important examples are $\log|x - y|$, $|x - y|^{\gamma-1}$ for some $\gamma > 0$ (although it is only singular for $0 < \gamma < 1$), and variants of them.

We introduce the idea of product integration by considering the special case of

$$\lambda u(x) - \int_a^b l(x, y) \log|y - x|\, u(y)\, dy = f(x), \quad a \le x \le b, \tag{12.5.2}$$

with the kernel

$$k(x, y) = l(x, y) \log|y - x|. \tag{12.5.3}$$

We assume that $l(x, y)$ is a well-behaved function (i.e. it is several times continuously differentiable), and initially, we assume the unknown solution $u(x)$ is also well-behaved. To solve (12.5.2), we define a method called the *product trapezoidal rule*.

Let $n \ge 1$ be an integer, $h = (b - a)/n$, and $x_j = a + jh$, $j = 0, 1, \ldots, n$. For general $u \in C[a, b]$, define

$$[l(x, y)u(y)]_n = \frac{1}{h}\left[(x_j - y)l(x, x_{j-1})u(x_{j-1}) + (y - x_{j-1})l(x, x_j)u(x_j)\right], \tag{12.5.4}$$

for $x_{j-1} \le y \le x_j$, $j = 1, \dots, n$ and $a \le x \le b$. This is piecewise linear in y, and it interpolates $l(x,y)u(y)$ at $y = x_0, \dots, x_n$, for all $x \in [a,b]$. Define a numerical approximation to the integral operator in (12.5.2) by

$$K_n u(x) \equiv \int_a^b [l(x,y)u(y)]_n \log|y-x|\, dy, \quad a \le x \le b. \tag{12.5.5}$$

This can also be written as

$$K_n u(x) = \sum_{j=0}^{n} w_j(x) l(x, x_j) u(x_j), \quad u \in C[a,b], \tag{12.5.6}$$

with weights

$$w_0(x) = \frac{1}{h} \int_{x_0}^{x_1} (x_1 - y) \log|x-y|\, dy, \tag{12.5.7}$$

$$w_n(x) = \frac{1}{h} \int_{x_{n-1}}^{x_n} (y - x_{n-1}) \log\|x-y|\, dy, \tag{12.5.8}$$

$$\begin{aligned} w_j(x) = {} & \frac{1}{h} \int_{x_{j-1}}^{x_j} (y - x_{j-1}) \log|x-y|\, dy \\ & + \frac{1}{h} \int_{x_j}^{x_{j+1}} (x_{j+1} - y) \log|x-y|\, dy, \quad j = 1, \dots, n-1. \end{aligned} \tag{12.5.9}$$

To approximate the integral equation (12.5.2), we use

$$\lambda u_n(x) - \sum_{j=0}^{n} w_j(x) l(x, x_j) u_n(x_j) = f(x), \quad a \le x \le b. \tag{12.5.10}$$

As with the Nyström method (12.4.4)–(12.4.6), this is equivalent to first solving the linear system

$$\lambda u_n(x_i) - \sum_{j=0}^{n} w_j(x_i) l(x_i, x_j) u_n(x_j) = f(x_i), \quad i = 0, \dots, n, \tag{12.5.11}$$

and then using the Nyström interpolation formula

$$u_n(x) = \frac{1}{\lambda} \left[f(x) + \sum_{j=0}^{n} w_j(x) l(x, x_j) u_n(x_j) \right], \quad a \le x \le b. \tag{12.5.12}$$

We leave it to the reader to check these assertions, since it is quite similar to what was done for the original Nyström method. With this method, we approximate those parts of the integrand in (12.5.2) that can be well-approximated by piecewise linear interpolation, and we integrate exactly the remaining more singular parts of the integrand.

Rather than using piecewise linear interpolation, other more accurate interpolation schemes could have been used to obtain a more rapidly convergent numerical method. Later in the section, we consider and illustrate the use of piecewise quadratic interpolation. We have also used evenly spaced node points $\{x_i\}$, but this is not necessary. The use of such evenly spaced nodes is an important case; but we will see later in the section that special choices of nonuniformly spaced node points are often needed for solving an integral equation such as (12.5.2).

Other singular kernel functions can be handled in a manner analogous to what has been done for (12.5.2). Consider the equation

$$\lambda u(x) - \int_a^b l(x,y)g(x,y)u(y)\,dy = f(x), \qquad a \le x \le b, \tag{12.5.13}$$

in which $g(x,y)$ is singular, with $l(x,y)$ and $u(x)$ as before. An important case is to take

$$g(x,y) = \frac{1}{|x-y|^{1-\gamma}}$$

for some $\gamma > 0$. To approximate (12.5.13), use the earlier approximation (12.5.4). Then

$$K_n u(x) = \int_a^b [l(x,y)u(y)]_n\, g(x,y)\,dy, \qquad a \le x \le b. \tag{12.5.14}$$

All arguments proceed exactly as before. To evaluate $K_n u(x)$, we need to evaluate the analogues of the weights in (12.5.7)–(12.5.9), where $\log|x-y|$ is replaced by $g(x,y)$. We assume these weights can be calculated in some practical manner, perhaps analytically. We consider further generalizations later in the section.

12.5.1 Error analysis

We consider the equation (12.5.13), with $l(x,y)$ assumed to be continuous. Further, we assume the following for $g(x,y)$:

$$c_g \equiv \sup_{a\le x\le b} \int_a^b |g(x,y)|\,dy < \infty, \tag{12.5.15}$$

$$\lim_{h\searrow 0} \omega_g(h) = 0, \tag{12.5.16}$$

where

$$\omega_g(h) \equiv \sup_{\substack{|x-\tau|\le h\\ a\le x,\tau\le b}} \int_a^b |g(x,y) - g(\tau,y)|\,dy.$$

These two properties can be shown to be true for both $\log|x-y|$ and $|x-y|^{\gamma-1}$, $\gamma > 0$. Such assumptions were used earlier in Subsection 2.8.1

in showing compactness of integral operators on $C[a,b]$, and we refer to that earlier material.

Theorem 12.5.1 *Assume the function $g(x,y)$ satisfies* (12.5.15)–(12.5.16)*, and assume $l(x,y)$ is continuous for $a \le x,y \le b$. For a given $f \in C[a,b]$, assume the integral equation*

$$\lambda u(x) - \int_a^b l(x,y)g(x,y)u(y)\,dy = f(x), \quad a \le x \le b,$$

is uniquely solvable. Consider the numerical approximation (12.5.14)*, with $[l(x,y)u(y)]_n$ defined with piecewise linear interpolation, as in* (12.5.4)*. Then for all sufficiently large n, say $n \ge N$, the equation* (12.5.14) *is uniquely solvable, and the inverse operators are uniformly bounded for such n. Moreover,*

$$\|u - u_n\|_\infty \le c\,\|Ku - K_n u\|_\infty, \quad n \ge N, \tag{12.5.17}$$

for suitable $c > 0$.

Proof. We can show that the operators $\{K_n\}$ of (12.5.14) are a collectively compact and pointwise convergent family on $C[a,b]$ to $C[a,b]$. This will prove the abstract assumptions **A1**–**A3** in Subsection 12.4.3; and by using Lemma 12.4.7, we can then apply Theorem 12.4.4. We note that **A1** is obvious from the definitions of K and K_n.

Let $\mathcal{S} = \{K_n v \mid n \ge 1 \text{ and } \|v\|_\infty \le 1\}$. For bounds on $\|K_n u\|_\infty$, first note that the piecewise linear interpolant z_n of a function $z \in C[a,b]$ satisfies

$$\|z_n\|_\infty \le \|z\|_\infty.$$

With this, it is straightforward to show

$$\|K_n u\|_\infty \le c_l c_g, \quad u \in C[a,b],\ \|u\|_\infty \le 1,$$

with

$$c_l \equiv \max_{a \le x,y \le b} |l(x,y)|.$$

This also shows the uniform boundedness of $\{K_n\}$, with

$$\|K_n\| \le c_l c_g, \quad n \ge 1.$$

For equicontinuity of $\mathcal{S}$, write

$$\begin{aligned} K_n u(x) - K_n u(\xi) &= \int_a^b [l(x,y)u(y)]_n \, g(x,y)\, dy \\ &\quad - \int_a^b [l(\xi,y)u(y)]_n \, g(\tau,y)\, dy \\ &= \int_a^b \{[l(x,y) - l(\xi,y)]\, u(y)\}_n \, g(x,y)\, dy \\ &\quad + \int_a^b [l(\xi,y)u(y)]_n \, [g(x,y) - g(\xi,y)]\, dy. \end{aligned}$$

This uses the linearity in z of the piecewise linear interpolation being used in defining $[z(y)]_n$. The assumptions on $g(x,y)$ and $l(x,y)$, together with $\|u\|_\infty \leq 1$, now imply

$$\left| \int_a^b \{[l(x,y) - l(\xi,y)]\, u(y)\}_n \, g(x,y) dy \right| \leq c_g \, \|u\|_\infty \max_{a \leq y \leq b} |l(x,y) - l(\xi,y)| \,.$$

Also,

$$\left| \int_a^b [l(\xi,y)u(y)]_n \, [g(x,y) - g(\xi,y)]\, dy \right| \leq c_l \, \|u\|_\infty \, \omega_g(|x - \xi|).$$

Combining these results shows the desired equicontinuity of $\mathcal{S}$, and it completes the proof of the abstract property **A3** needed in applying the collectively compact operator framework.

We leave the proof of **A2** as an exercise for the reader. To complete the proof of the theorem, we apply Lemma 12.4.7 and Theorem 12.4.4. The constant c is the uniform bound on $\|(\lambda - K_n)^{-1}\|$ for $n \geq N$. □

Example 12.5.2 Let $z_n(y)$ denote the piecewise linear interpolant of $z(y)$, as used above in defining the product trapezoidal rule. It is a well-known standard result that

$$|z(y) - z_n(y)| \leq \frac{h^2}{8} \, \|z''\|_\infty, \quad z \in C^2[a,b].$$

Thus if $l(x,\cdot) \in C^2[a,b]$, $a \leq x \leq b$, and if $u \in C^2[a,b]$, then (12.5.17) implies

$$\|u - u_n\|_\infty \leq \frac{ch^2}{8} \max_{a \leq x,y \leq b} \left| \frac{\partial^2 l(x,y)u(y)}{\partial y^2} \right|, \quad n \geq N. \tag{12.5.18}$$

The above ideas for solving (12.5.13) will generalize easily to higher degrees of piecewise polynomial interpolation. All elements of the above proof

also generalize, and we obtain a theorem analogous to Theorem 12.5.1. In particular, suppose $[l(\tau, y)u(y)]_n$ is defined using piecewise polynomial interpolation of degree $m \geq 0$. Assume $l(x, \cdot) \in C^{m+1}[a, b]$, $a \leq x \leq b$, and $u \in C^{m+1}[a, b]$. Then

$$\|u - u_n\|_\infty \leq ch^{m+1} \max_{a \leq x, y \leq b} \left| \frac{\partial^{m+1} l(x, y)u(y)}{\partial y^{m+1}} \right|, \quad n \geq N, \tag{12.5.19}$$

for a suitable constant $c > 0$. When using piecewise quadratic interpolation, the method (12.5.14) is called the *product Simpson rule*; and according to (12.5.19), its rate of convergence is at least $\mathcal{O}(h^3)$. □

12.5.2 Generalizations to other kernel functions

Many singular integral equations are not easily written in the form (12.5.13) with a function $l(x, y)$ which is smooth and a function $g(x, y)$ for which weights such as those in (12.5.7)–(12.5.9) can be easily calculated. For such equations, we assume instead that the singular kernel function $k(x, y)$ can be written in the form

$$k(x, y) = \sum_{j=1}^{r} l_j(x, y) g_j(x, y) \tag{12.5.20}$$

with each $l_j(x, y)$ and $g_j(x, y)$ satisfying the properties listed above for $l(x, y)$ and $g(x, y)$. We now have an integral operator written as a sum of integral operators of the form used in (12.5.13):

$$Ku(x) = \sum_{j=1}^{r} K_j u(x) = \sum_{j=1}^{r} \int_a^b l_j(x, y) g_j(x, y) u(y)\, dy, \quad u \in C[a, b].$$

Example 12.5.3 Consider the integral equation

$$u(x) - \int_0^\pi u(y) \log |\cos x - \cos y|\, dy = 1, \quad 0 \leq x \leq \pi. \tag{12.5.21}$$

One possibility for the kernel function $k(x, y) = \log |\cos x - \cos y|$ is to write

$$k(x, y) = \underbrace{|x - y|^{1/2} \log |\cos x - \cos y|}_{=l(x,y)} \underbrace{|x - y|^{-1/2}}_{=g(x,y)}.$$

Unfortunately, this choice of $l(x, y)$ is continuous without being differentiable; and the function $l(x, y)$ needs to be differentiable in order to have the numerical method converge with sufficient speed. A better choice is to

	Product trapezoidal		*Product Simpson*	
n	$\|u-u_n\|_\infty$	*Ratio*	$\|u-u_n\|_\infty$	*Ratio*
2	9.50E−3		2.14E−4	
4	2.49E−3	3.8	1.65E−5	13.0
8	6.32E−4	3.9	1.13E−6	14.6
16	1.59E−4	4.0	7.25E−8	15.6
32	3.98E−5	4.0	4.56E−9	15.9

TABLE 12.3. Product trapezoidal and product Simpson examples for (12.5.21)

use

$$\begin{aligned} k(x,y) &= \log\left|2\sin\frac{1}{2}(x-y)\,\sin\frac{1}{2}(x+y)\right| \\ &= \log\left[\frac{2\sin\frac{1}{2}(x-y)\,\sin\frac{1}{2}(x+y)}{(x-y)(x+y)(2\pi-x-y)}\right] + \log|x-y| \\ &\quad + \log(x+y) + \log(2\pi-x-y). \end{aligned} \tag{12.5.22}$$

This is of the form (12.5.20) with $g_1 = l_2 = l_3 = l_4 \equiv 1$ and

$$\begin{aligned} l_1(x,y) &= \log\left[\frac{2\sin\frac{1}{2}(x-y)\,\sin\frac{1}{2}(x+y)}{(x-y)(x+y)(2\pi-x-y)}\right], \\ g_2(x,y) &= \log|x-y|\,, \\ g_3(x,y) &= \log(x+y), \\ g_4(x,y) &= \log(2\pi-x-y). \end{aligned}$$

The function $l_1(x,y)$ is infinitely differentiable on $[0,2\pi]$; and the functions g_2, g_3, and g_4 are singular functions for which the needed integration weights are easily calculated.

We solve (12.5.21) with both the product trapezoidal rule and the product Simpson rule, and error results are given in Table 12.3. The decomposition (12.5.22) is used to define the approximating operators. With the operator with kernel $l_1(x,y)g_1(x,y)$, we use the regular Simpson rule. The true solution of the equation is

$$u(x) \equiv \frac{1}{1+\pi\log 2} \doteq 0.31470429802.$$

Note that the error for the product trapezoidal rule is consistent with (12.5.18). But for the product Simpson rule, we appear to have an error behaviour of $\mathcal{O}(h^4)$, whereas that predicted by (12.5.19) is only $\mathcal{O}(h^3)$. This is discussed further below. □

12.5.3 Improved error results for special kernels

If we consider again the error formula (12.5.17), the error result (12.5.19) was based on applying standard error bounds for polynomial interpolation to bounding the numerical integration error $\|Ku - K_n u\|_\infty$. We know that for many ordinary integration rules (e.g. Simpson's rule), there is an improvement in the speed of convergence over that predicted by the polynomial interpolation error, and this improvement is made possible by fortuitous cancellation of errors when integrating. Thus it is not surprising that the same type of cancellation occurs with the error $\|Ku - K_n u\|_\infty$ in product Simpson integration, as is illustrated in Table 12.3.

For the special cases of $g(x, y)$ equal to $\log|x-y|$ and $|x-y|^{\gamma-1}$, deHoog and Weiss [66] improved on the bound (12.5.19). In [66], they first extended known asymptotic error formulas for ordinary composite integration rules to product integration formulas; and then these results were further extended to estimate $Ku - K_n u$ for product integration methods of solving singular integral equations. For the case of the product Simpson's rule, their results state that if $u \in C^4[a, b]$, then

$$\|Ku - K_n u\|_\infty \leq \begin{cases} ch^4|\log h|, & g(x,y) = \log|x-y|, \\ ch^{3+\gamma}, & g(x,y) = |x-y|^{\gamma-1}. \end{cases} \tag{12.5.23}$$

This is in agreement with the results in Table 12.3.

12.5.4 Product integration with graded meshes

The rate of convergence results (12.5.19) and (12.5.23) both assume that the unknown solution $u(x)$ possesses several continuous derivatives. In fact, $u(x)$ seldom is smoothly differentiable, but rather has somewhat singular behaviour in the neighborhood of the endpoints of the interval $[a, b]$ on which the integral equation is being solved. In the following, this is made more precise; and we also give a numerical method which restores the speed of convergence seen above with smoothly differentiable unknown solution functions.

To examine the differentiability of the solution $u(x)$ of a general integral equation $(\lambda - K)u = f$, the differentiability of the kernel function $k(x, y)$ allows the smoothness of $f(x)$ to be carried over to that of $u(x)$: Use

$$\frac{d^j u(x)}{dx^j} = \frac{1}{\lambda}\left[\frac{d^j f(x)}{dx^j} + \int_a^b \frac{\partial^j k(x,y)}{\partial x^j} u(y)dy\right].$$

But if the kernel function is not differentiable, then the integral operator need not be smoothing. To see that the integral operator K with kernel $\log|x - y|$ is not smoothing in the manner that is true with differentiable

kernel functions, let $u_0(x) \equiv 1$ on the interval $[0,1]$, and calculate $Ku_0(x)$:

$$Ku_0(x) = \int_0^1 \log|x-y|\,dy = x\log x + (1-x)\log(1-x) - 1, \quad 0 \le x \le 1. \tag{12.5.24}$$

The function $Ku_0(x)$ is not continuously differentiable on $[0,1]$, whereas the function $u_0(x)$ is a C^∞ function. This formula also contains the typical type of singular behaviour that appears in the solution when solving a second kind integral equation with a kernel function $k(x,y) = l(x,y)\log|x-y|$.

We give the main result of Schneider [202] on the regularity behaviour of solutions of $(\lambda - K)\,u = f$ for special weakly singular kernel functions. As notation, introduce the following spaces:

$$C^{(0,\beta)}[a,b] = \left\{ g \in C[a,b] \;\middle|\; d_\beta(g) \equiv \sup_{a\le x,\xi\le b} \frac{|g(x)-g(\xi)|}{|x-\xi|^\beta} < \infty \right\} \tag{12.5.25}$$

for $0 < \beta < 1$, and

$$C^{(0,1)}[a,b] = \left\{ g \in C[a,b] \;\middle|\; \sup_{a\le x,\xi\le b} \frac{|g(x)-g(\xi)|}{|x-\xi|\log|B/(x-\xi)|} < \infty \right\},$$

for some $B > b-a$. For $0 < \beta < 1$, $C^{(0,\beta)}[a,b]$ are the standard Hölder spaces introduced in Subsection 1.4.1 of Chapter 1.

Theorem 12.5.4 *Let $k \ge 0$ be an integer, and let $0 < \gamma \le 1$. Assume $f \in C^{(0,\gamma)}[a,b]$, $f \in C^k(a,b)$, and*

$$(x-a)^i(b-x)^i f^{(i)}(x) \in C^{(0,\gamma)}[a,b], \quad i = 1,\dots,k.$$

Also assume $L \in C^{k+1}(D)$ with $D = [a,b]\times[a,b]$. Finally, assume the integral equation

$$\lambda u(x) - \int_a^b l(x,y) g_\gamma(x-y) u(y)\,dy = f(x), \quad a \le x \le b \tag{12.5.26}$$

with

$$g_\gamma(u) \equiv \begin{cases} u^{\gamma-1}, & 0 < \gamma < 1, \\ \log|u|, & \gamma = 1 \end{cases}$$

is uniquely solvable. Then

(a) *The solution $u(x)$ satisfies $u \in C^{(0,\gamma)}[a,b]$, $u \in C^k(a,b)$, and*

$$u_i(x) \equiv (x-a)^i(b-x)^i u^{(i)}(x) \in C^{(0,\gamma)}[a,b], \quad i = 1,\dots,k. \tag{12.5.27}$$

Further, $u_i(a) = u_i(b) = 0$, $i = 1,\dots,k$.

(b) *For* $0 < \gamma < 1$,

$$\left|u^{(i)}(x)\right| \le c_i(x-a)^{\gamma-i}, \quad a < x \le \frac{1}{2}(a+b), \quad i = 1, \ldots, k. \tag{12.5.28}$$

With $\gamma = 1$, *for any* $\epsilon \in (0, 1)$,

$$\left|u^{(i)}(x)\right| \le c_i(x-a)^{1-\epsilon-i}, \quad a < x \le \frac{1}{2}(a+b), \quad i = 1, \ldots, k, \tag{12.5.29}$$

with c_i *dependent on* ϵ. *Analogous results are true for* x *in a neighborhood of* b, *with* $x - a$ *replaced by* $b - x$.

A proof of this theorem is given in [202, p. 63]. In addition, more detail on the asymptotic behaviour of $u(x)$ for x near to either a or b is given in the same reference and in Graham [97], bringing in functions of the type seen on the right side of (12.5.24) for the case of logarithmic kernel functions.

This theorem says we should expect endpoint singularities in $u(x)$ of the form $(x-a)^\gamma$ and $(b-x)^\gamma$ for the case $g(x, y) = |x-y|^{\gamma-1}$, $0 < \gamma < 1$, with corresponding results for the logarithmic kernel. Thus the approximation of the unknown $u(x)$ should be based on such behaviour. We do so by introducing the concept of a *graded mesh*, an idea developed in Rice [194] for the types of singular functions considered here.

We first develop the idea of a graded mesh for functions on $[0, 1]$ with the singular behaviour in the function occurring at 0; and then the construction is extended to other situations by a simple change of variables. The singular behaviour in which we are interested is $u(x) = x^\gamma$, $\gamma > 0$. For a given integer $n \ge 1$, define

$$x_j = \left(\frac{j}{n}\right)^q, \quad j = 0, 1, \ldots, n, \tag{12.5.30}$$

with the real number $q \ge 1$ to be specified later. For $q > 1$, this is an example of a *graded mesh*, and it is the one introduced and studied in Rice [194]. For a given integer $m \ge 0$, let a partition of $[0, 1]$ be given:

$$0 \le \mu_0 < \cdots < \mu_m \le 1. \tag{12.5.31}$$

Define interpolation nodes on each subinterval $[x_{j-1}, x_j]$ by

$$x_{ji} = x_{j-1} + \mu_i h_j, \quad i = 0, 1, \ldots, m, \quad h_j \equiv x_j - x_{j-1}.$$

Let $P_n u(x)$ be the piecewise polynomial function which is of degree $\le m$ on each subinterval $[x_{j-1}, x_j]$ and which interpolates $u(x)$ at the nodes $\{x_{j0}, \ldots, x_{jm}\}$ on that subinterval. To be more explicit, let

$$L_i(\mu) = \prod_{\substack{k=0 \\ k \ne i}}^{m} \frac{\mu - \mu_k}{\mu_i - \mu_k}, \quad i = 0, 1, \ldots, m$$

which are the basis functions associated with interpolation at the nodes of (12.5.31). Then

$$P_n u(x) = \sum_{i=0}^{m} L_i\left(\frac{x - x_{j-1}}{h_j}\right) u(x_{ji}), \quad x_{j-1} \le x \le x_j, \quad j = 1, \ldots, n, \tag{12.5.32}$$

If $\mu_0 > 0$ or $\mu_m < 1$, then $P_n u(x)$ is likely to be discontinuous at the interior breakpoints $x_1, \ldots, x_{n-1}$. We now present the main result from Rice [194].

Lemma 12.5.5 *Let n, m, $\{x_j\}$, $\{x_{ji}\}$, and P_n be as given in the preceding paragraph. For $0 < \gamma < 1$, assume $u \in C^{(0,\gamma)}[0,1] \cap C^{m+1}(0,1]$, with*

$$\left|u^{(m+1)}(x)\right| \le c_{\gamma,m}(u) x^{\gamma-(m+1)}, \quad 0 < x \le 1. \tag{12.5.33}$$

Then for

$$q \ge \frac{m+1}{\gamma}, \tag{12.5.34}$$

we have,

$$\|u - P_n u\|_\infty \le \frac{c}{n^{m+1}}, \tag{12.5.35}$$

with c a constant independent of n. For $1 \le p < \infty$, let

$$q > \frac{p(m+1)}{1 + p\gamma}. \tag{12.5.36}$$

Then

$$\|u - P_n u\|_p \le \frac{c}{n^{m+1}}, \tag{12.5.37}$$

with $\|\cdot\|_p$ denoting the standard p-norm for $L^p(0,1)$.

A proof of the result can be found in [18, p. 128]. In the language of Rice [194], a function $u(x)$ satisfying the conditions stated in Lemma 12.5.5 is said to be of *Type*$(\gamma, m+1)$.

The earlier product integration methods were based on using interpolation on a uniform subdivision of the interval $[a,b]$. Now we use the same form of interpolation, but base it on a graded mesh for $[a,b]$. Given an even $n \ge 2$, define

$$x_j = a + \left(\frac{2j}{n}\right)^q \left(\frac{b-a}{2}\right), \quad x_{n-j} = b + a - x_j, \ j = 0, 1, \ldots, \frac{n}{2}.$$

Use the partition (12.5.31) as the basis for polynomial interpolation of degree m on each of the intervals $[x_{j-1}, x_j]$, for $j = 1, \ldots, \frac{1}{2}n$, just as was done in (12.5.32); and use the partition

$$0 \le 1 - \mu_m < \cdots < 1 - \mu_0 \le 1$$

when defining the interpolation on the subintervals $[x_{j-1}, x_j]$ of the remaining half $[\frac{1}{2}(a+b), b]$. In the integral equation (12.5.26), replace $[l(x,y)u(y)]$ with $[l(x,y)u(y)]_n$ using the interpolation just described. For the resulting approximation

$$\lambda u_n(x) - \int_a^b [l(x,y)u_n(y)]_n g_\gamma(x-y)\,dy = f(x), \quad a \le x \le b, \tag{12.5.38}$$

we have the following convergence result.

Theorem 12.5.6 *Consider again the integral equation* (12.5.26)*, and assume the same assumptions as for Theorem 12.5.4, but with the integer k replaced by $m+1$, where m is the integer used in defining the interpolation of the preceding paragraph. Then the approximating equation* (12.5.38) *is uniquely solvable for all sufficiently large n, say $n \ge N$, and the inverse operator for the equation is uniformly bounded for $n \ge N$. If $0 < \gamma < 1$, then choose the grading exponent q to satisfy*

$$q \ge \frac{m+1}{\gamma}. \tag{12.5.39}$$

If $\gamma = 1$, then choose

$$q > m+1. \tag{12.5.40}$$

With such choices, the approximate solution u_n satisfies

$$\|u - u_n\|_\infty \le \frac{c}{n^{m+1}}. \tag{12.5.41}$$

Proof. The proof is a straightforward generalization of the method of proof used in Theorem 12.5.1, resulting in the error bound

$$\|u - u_n\|_\infty \le c\,\|Ku - K_n u\|_\infty, \quad n \ge N.$$

Combine Theorem 12.5.4 and Lemma 12.5.5 to complete the proof.

This theorem is from Schneider [203, Theorem 2], and he also allows for greater generality in the singularity in $u(x)$ than has been assumed here. In addition, he extends results of deHoog and Weiss [66], such as (12.5.23), to the use of graded meshes. □

Graded meshes are used with other problems in which there is some kind of singular behaviour in the functions being considered. For example, they are used in solving boundary integral equations for the planar Laplace's equation for regions whose boundary has corners.

12.5.5 The relationship of product integration and collocation methods

Recall the earlier discussion of the collocation method in Section 12.1 and Section 12.3. It turns out that collocation methods can be regarded as

product integration methods, and occasionally there is an advantage to doing so.

Recalling this earlier discussion, let P_n be the interpolatory projection operator from $C(D)$ onto the interpolatory approximating space V_n. Then the collocation solution of $(\lambda - K)\, u = f$ can be regarded abstractly as $(\lambda - P_n K)\, u_n = P_n f$, and the iterated collocation solution

$$\widehat{u}_n = \frac{1}{\lambda}\,(f + K u_n)$$

is the solution of the equation

$$(\lambda - K P_n)\widehat{u}_n = f. \tag{12.5.42}$$

Define a numerical integral operator by

$$K_n u(x) = K P_n u(x) = \int_D k(x,y)(P_n u)(y)\, dy. \tag{12.5.43}$$

This is product integration with $l(x,y) \equiv 1$ and $g(x,y) = k(x,y)$. Thus the iterated collocation solution $\widehat{u}_n$ of (12.5.42) is simply the Nyström solution when defining K_n using the simple product integration formula (12.5.43). Since the collocation solution satisfies $u_n = P_n \widehat{u}_n$, we can use results from the error analysis of product integration methods to analyze collocation methods.

Exercise 12.5.1 Prove the property **A2** in the proof of Theorem 12.5.1.

Exercise 12.5.2 Develop a practical product trapezoidal rule (with even spacing) for the numerical solution of

$$\lambda u(x) - \int_0^\pi u(y) \log|\sin(x-y)|\, dy = f(x), \quad 0 \le x \le \pi,$$

assuming λ is so chosen that the integral equation is uniquely solvable. Program your procedure. Solve the equation approximately with $f(x) = 1$ and $f(x) = e^{\sin x}$. Do numerical examples with varying values of n, as in Example 12.5.3.

Exercise 12.5.3 Develop a product integration Nyström method for solving

$$\lambda u(x) - \int_0^1 \frac{c\,\ell(x,y)u(y)}{c^2 + (x-y)^2}\, dy = f(x), \quad 0 \le x \le 1$$

where c is a very small positive number. Assume $\ell(x,y)$ and its low-order derivatives are "well-behaved" functions; and note that the above kernel function is very peaked for small values of c. Define the numerical integration operators, and discuss the error in them as approximations to the original operator. Discuss convergence of your Nyström method.

Exercise 12.5.4 Consider numerically approximating

$$I = \int_0^1 x^\alpha dx, \quad 0 < \alpha < 1$$

using the trapezoidal rule with a graded mesh of the form (12.5.30). How should the grading parameter q be chosen so as to insure that the rate of convergence is $\mathcal{O}(n^{-2})$?
Hint: Consider the error on each subinterval $[x_{i-1}, x_i]$, and consider separately the cases of $i = 1$ and $i > 1$. Choose q to make the error on $[x_0, x_1]$ of size $\mathcal{O}(n^{-2})$. Then examine the error on the remaining subintervals and the total on $[x_1, 1]$. Recall the use of integrals to approximate summations.

12.6 Iteration methods

In this chapter we have given a number of ways to solve approximately the second kind integral equation

$$\lambda u(x) - \int_D k(x,y)u(y)\,dy = f(x), \quad x \in D, \tag{12.6.1}$$

or symbolically,

$$(\lambda - K)\,u = f. \tag{12.6.2}$$

Each numerical method results in a sequence of approximating equations

$$(\lambda - K_n)\,u_n = f_n, \qquad n \geq 1. \tag{12.6.3}$$

The operator K_n can be a degenerate kernel integral operator, an approximation based on an orthogonal or interpolatory projection, a numerical integration operator, or some other approximation not considered here. In each case, the functional equation (12.6.3) reduces to an equivalent finite linear system,

$$A_n \boldsymbol{u}_n = \boldsymbol{f}_n. \tag{12.6.4}$$

Let q_n denote the order of this system. Then solving the system directly by Gaussian elimination will cost approximately $\frac{2}{3}q_n^3$ arithmetic operations. If q_n is quite large, then other methods of solution must be found, and such methods are generally some type of iterative procedure. In this section we begin by describing a general framework for developing iteration methods. We follow it with an iteration method for solving the linear system (12.4.5) associated with the Nyström method for solving (12.6.1).

A general method for developing an iteration method for solving linear systems, such as (12.6.4), is as follows. Denote an initial guess for $\boldsymbol{u}_n$ by $\boldsymbol{u}_n^{(0)}$, and calculate the residual

$$\boldsymbol{r}_n^{(0)} = \boldsymbol{f}_n - A_n \boldsymbol{u}_n^{(0)}.$$

Then we have

$$A_n\left(\boldsymbol{u}_n - \boldsymbol{u}_n^{(0)}\right) = \boldsymbol{r}_n^{(0)}$$

and

$$\boldsymbol{u}_n = \boldsymbol{u}_n^{(0)} + A_n^{-1}\boldsymbol{r}_n^{(0)}.$$

Let C_n denote an approximation of the inverse matrix A_n^{-1}. We obtain an improvement on $\boldsymbol{u}_n^{(0)}$ by using

$$\boldsymbol{u}_n^{(1)} = \boldsymbol{u}_n^{(0)} + C_n\boldsymbol{r}_n^{(0)}.$$

Repeating this process leads to the iteration method

$$\boldsymbol{r}_n^{(\kappa)} = \boldsymbol{f}_n - A_n\boldsymbol{u}_n^{(\kappa)}, \tag{12.6.5}$$

$$\boldsymbol{u}_n^{(\kappa+1)} = \boldsymbol{u}_n^{(\kappa)} + C_n\boldsymbol{r}_n^{(\kappa)}, \quad \kappa = 0, 1, 2, \ldots. \tag{12.6.6}$$

This is called *iterative improvement* or the *residual correction method*. Many iteration methods for solving linear systems can be defined within this schema. We can use the same framework to develop iteration methods for solving the operator equation (12.6.3).

How do we obtain C_n? Since the linear system (12.6.4) is equivalent to the operator equation in (12.6.3), and since these approximating equations are inter-related as approximations to (12.6.2), we use information associated with $(\lambda - K_m)\, u_m = f_m$ and $A_m\boldsymbol{u}_m = \boldsymbol{f}_m$, $m < n$, to develop implicitly an approximation $C_n \approx A_n^{-1}$. When only a single value m is used, usually with m much smaller than n, we call it a *two-grid iteration method*. The construction of C_n varies with the way in which K_n is defined.

12.6.1 A two-grid iteration method for the Nyström method

Use the Nyström scheme of Section 12.4 as our numerical method (12.6.3). For the linear system $A_n\boldsymbol{u}_n = \boldsymbol{f}_n$, recall that

$$(A_n)_{i,j} = \lambda\delta_{i,j} - w_j k(x_i, x_j), \tag{12.6.7}$$

$$(\boldsymbol{f}_n)_i = f(x_i) \tag{12.6.8}$$

for $1 \le i, j \le q_n$. We first develop an iteration method for the operator equation $(\lambda - K_n)\, u_n = f$, and then we specialize it to the linear system $A_n\boldsymbol{u}_n = \boldsymbol{f}_n$.

Begin by using the framework of the residual correction method as applied to the operator equation $(\lambda - K_n)\, u_n = f$. Assume $u_n^{(0)}$ is an initial estimate of the solution u_n. Define the residual

$$r_n^{(0)} = f - (\lambda - K_n)u_n^{(0)}. \tag{12.6.9}$$

This leads to

$$(\lambda - K_n)\left[u_n - u_n^{(0)}\right] = r_n^{(0)}$$

and so

$$u_n = u_n^{(0)} + (\lambda - K_n)^{-1} r_n^{(0)}. \tag{12.6.10}$$

By estimating $(\lambda - K_n)^{-1} r_n^{(0)}$, we can define an iteration method for solving $(\lambda - K_n)\, u_n = f$.

For some $m < n$, assume we can solve directly the approximating equation $(\lambda - K_m)\, w = z$, for arbitrary $z \in C(D)$. Then consider the approximation

$$(\lambda - K_n)^{-1} r_n^{(0)} \approx (\lambda - K_m)^{-1} r_n^{(0)}. \tag{12.6.11}$$

Using it in (12.6.10), define

$$u_n^{(1)} = u_n^{(0)} + (\lambda - K_m)^{-1} r_n^{(0)}.$$

The general iteration is defined by

$$r_n^{(\kappa)} = y - (\lambda - K_n) u_n^{(\kappa)} \tag{12.6.12}$$

$$u_n^{(\kappa+1)} = u_n^{(\kappa)} + (\lambda - K_m)^{-1} r_n^{(\kappa)}, \quad \kappa = 0, 1, 2, \ldots. \tag{12.6.13}$$

As notation, we call $(\lambda - K_m)\, u_m = f$ the *coarse mesh approximation* and $(\lambda - K_n)\, u_n = f$ the *fine mesh approximation*. In (12.6.13), we solve the coarse mesh equation

$$(\lambda - K_m)\, \delta_{n,m}^{(\kappa)} = r_n^{(\kappa)} \tag{12.6.14}$$

and then define

$$u_n^{(\kappa+1)} = u_n^{(\kappa)} + \delta_{n,m}^{(\kappa)} \tag{12.6.15}$$

as a new fine mesh approximation.

This iteration turns out to be less than ideal, but it does converge when m is chosen sufficiently large: $u_n^{(\kappa)} \to u_n$ as $\kappa \to \infty$, and the speed of convergence is uniform with respect to $n > m$. Rather than analyzing this iteration method, we turn to another method that is better in practice.

In some early work on iteration methods, it was thought that the iteration method (12.6.12)–(12.6.13) would not converge in most cases. This was incorrect; but the method designed to replace it is a significantly better iteration method in most situations.

Recall the equations (12.6.9)–(12.6.10) for relating the residual and the error in an initial guess $u_n^{(0)}$ for solving the approximating equation $(\lambda - K_n)\, u_n = y$. Rather than approximating the error $e_n \equiv u_n - u_n^{(0)}$, which may be anywhere in some neighborhood of the zero element in the function space $C(D)$, we introduce a new unknown which can vary over only a much smaller (and compact) neighborhood of the origin. This turns out to be an aid in the accurate approximation of e_n.

Since

$$(\lambda - K_n)\, e_n = r_n^{(0)} \equiv f - (\lambda - K_n)\, u_n^{(0)}, \tag{12.6.16}$$

we can solve for e_n to obtain

$$e_n = \frac{1}{\lambda}[r_n^{(0)} + K_n e_n].$$

Introduce the new unknown

$$\delta_n = K_n e_n \tag{12.6.17}$$

so that

$$e_n = \frac{1}{\lambda}[r_n^{(0)} + \delta_n]. \tag{12.6.18}$$

Substitute this into (12.6.16) and simplify, obtaining

$$(\lambda - K_n)\,\delta_n = K_n r_n^{(0)}. \tag{12.6.19}$$

We estimate the unknown δ_n by solving the coarse mesh equation

$$(\lambda - K_m)\,\delta_n^{(0)} = K_n r_n^{(0)}.$$

Then define

$$e_n^{(0)} = \frac{1}{\lambda}[r_n^{(0)} + \delta_n^{(0)}],$$
$$u_n^{(1)} = u_n^{(0)} + \frac{1}{\lambda}[r_n^{(0)} + \delta_n^{(0)}].$$

The general iteration is defined by

$$r_n^{(\kappa)} = f - (\lambda - K_n)u_n^{(\kappa)}, \tag{12.6.20}$$
$$(\lambda - K_m)\delta_n^{(\kappa)} = K_n r_n^{(\kappa)}, \tag{12.6.21}$$
$$u_n^{(\kappa+1)} = u_n^{(\kappa)} + \frac{1}{\lambda}[r_n^{(\kappa)} + \delta_n^{(\kappa)}] \tag{12.6.22}$$

for $\kappa = 0, 1, \ldots$. This iteration method is usually superior to that given in (12.6.12)–(12.6.13). When the phrase "two-grid iteration for Nyström's method" is used, it generally refers to the iteration method (12.6.20)–(12.6.22).

To have this fit within the framework of the residual correction method, it is straightforward to show

$$u_n^{(\kappa+1)} = u_n^{(\kappa)} + \mathcal{C}_n r_n^{(\kappa)}, \tag{12.6.23}$$
$$\mathcal{C}_n = \frac{1}{\lambda}\left[I + (\lambda - K_m)^{-1}K_n\right]. \tag{12.6.24}$$

From the identity

$$(\lambda - K_n)^{-1} = \frac{1}{\lambda}\left[I + (\lambda - K_n)^{-1}K_n\right],$$

we obtain

$$(\lambda - K_n)^{-1} \approx \frac{1}{\lambda}\left[I + (\lambda - K_m)^{-1}K_n\right] \equiv \mathcal{C}_n.$$

12.6.2 Convergence analysis

As preliminaries to the convergence analysis of the iteration method (12.6.20)–(12.6.22), return to Section 12.4 and the error analysis given there. Let $N(\lambda)$ be so chosen that

$$\|(K_p - K)\, K_p\| \leq \frac{|\lambda|}{2\left\|(\lambda - K)^{-1}\right\|} \quad \text{for } p \geq N(\lambda).$$

Then Theorem 12.4.4 implies that for $p \geq N(\lambda)$,

$$\begin{aligned} \left\|(\lambda - K_p)^{-1}\right\| &\leq \frac{1 + \left\|(\lambda - K)^{-1}\right\| \|K_p\|}{|\lambda| - \|(\lambda - K)^{-1}\| \, \|(K - K_p)K_p\|} \\ &\leq \frac{2}{|\lambda|}\left[1 + \left\|(\lambda - K)^{-1}\right\| \|K_p\|\right], \end{aligned} \tag{12.6.25}$$

which is uniformly bounded with respect to p. In fact, (12.4.11) gives a uniform bound for $\|K_p\|$ for $p \geq 1$, and thus the right side of (12.6.25) is also bounded. For later reference, let

$$B_I(\lambda) = \sup_{p \geq N(\lambda)} \left\|(\lambda - K_p)^{-1}\right\|.$$

Theorem 12.6.1 *Assume the integral equation $(\lambda - K)\, u = f$ is uniquely solvable for all $f \in C(D)$, and let $k(x, y)$ be continuous for $x, y \in D$. Assume the numerical integration scheme*

$$\int_D g(t)\, dt \approx \sum_{j=1}^{q_n} w_{n,j} g(t_{n,j}), \qquad n \geq 1 \tag{12.6.26}$$

is convergent as $n \to \infty$, for all $g \in C(D)$. Then if m is chosen sufficiently large, the iteration method (12.6.20)–(12.6.22) *is convergent, i.e.*

$$u_n^{(\kappa)} \to u_n \quad \text{as } \kappa \to \infty$$

for all $n > m$.

Proof. In the iteration (12.6.20)–(12.6.22), we restrict $m \geq N(\lambda)$ and $n > m$. We are interested in the error $u_n - u_n^{(\kappa)}$. From the definition,

$$\begin{aligned}
u_n - u_n^{(\kappa+1)} &= u_n - u_n^{(\kappa)} - \frac{1}{\lambda}[r_n^{(\kappa)} + \delta_n^{(\kappa)}] \\
&= u_n - u_n^{(\kappa)} - \frac{1}{\lambda}[r_n^{(\kappa)} + (\lambda - K_m)^{-1} K_n r_n^{(\kappa)}] \\
&= u_n - u_n^{(\kappa)} - \frac{1}{\lambda}\left[I + (\lambda - K_m)^{-1} K_n\right] r_n^{(\kappa)} \\
&= u_n - u_n^{(\kappa)} - \frac{1}{\lambda}\left[I + (\lambda - K_m)^{-1} K_n\right](\lambda - K_n)\left[u_n - u_n^{(\kappa)}\right] \\
&= \left\{I - \frac{1}{\lambda}\left[I + (\lambda - K_m)^{-1} K_n\right](\lambda - K_n)\right\}\left[u_n - u_n^{(\kappa)}\right] \\
&= \frac{1}{\lambda}(\lambda - K_m)^{-1}(K_n - K_m) K_n \left[u_n - u_n^{(\kappa)}\right].
\end{aligned}$$

Thus

$$u_n - u_n^{(\kappa+1)} = \mathcal{M}_{m,n}\left[u_n - u_n^{(\kappa)}\right] \tag{12.6.27}$$

with

$$\mathcal{M}_{m,n} = \frac{1}{\lambda}(\lambda - K_m)^{-1}(K_n - K_m) K_n. \tag{12.6.28}$$

To show convergence of $u_n^{(\kappa)}$ to u_n as $\kappa \to \infty$, we need to examine the size of $\mathcal{M}_{m,n}$.

Introduce a set

$$\Psi = \{K_p v \mid p \geq 1 \text{ and } \|v\|_\infty \leq 1\}. \tag{12.6.29}$$

From **A1**-**A3** in Subsection 12.4.3 and Lemma 12.4.7, the set Ψ has compact closure in $C(D)$. Define

$$\begin{aligned}
B_K &= \sup_{p \geq 1} \|K_p\|, \\
a_m &= \sup_{p \geq m} \sup_{l \geq 1} \|(K - K_p) K_l\|.
\end{aligned}$$

We have

$$a_m \leq \sup_{p \geq m} \sup_{z \in \Psi} \|Kz - K_p z\|_\infty.$$

Since Ψ has compact closure from **A3**, and since $\{K_p\}$ is pointwise convergent on $C(D)$ from **A2**, it is relatively straightforward to show (a) the constant B_K is finite, and (b)

$$a_m \to 0 \quad \text{as } m \to \infty \tag{12.6.30}$$

monotonically.

For our quadrature operators $\{K_p \mid p \geq 1\}$, B_K is bounded from (12.4.11); and from the constructions used in Lemma 12.4.2,

$$\|(K - K_p)K_l\| = \max_{x \in D} \sum_{j=1}^{q_l} |w_{l,j} e_p(x, x_{l,j})| .$$

In this formula, $e_p(t, s)$ is the numerical integration error

$$e_p(x, y) = \int_D k(x, v)k(v, y)\, dv - \sum_{j=1}^{q_p} w_j k(x, x_j)k(x_j, y), \quad x, y \in D,\ p \geq 1.$$

It was shown in Lemma 12.4.2 to converge to zero uniformly for $x, y \in D$.

From (12.6.28), for $n \geq m$,

$$\begin{aligned}
\|\mathcal{M}_{m,n}\| &\leq \frac{1}{|\lambda|} \left\|(\lambda - K_m)^{-1}\right\| \|(K_n - K_m)K_n\| \\
&= \frac{1}{|\lambda|} \left\|(\lambda - K_m)^{-1}\right\| \|[(K - K_m) - (K - K_n)]\, K_n\| \\
&\leq \frac{1}{|\lambda|} B_I(\lambda)(a_m + a_n) \\
&\leq \frac{2a_m}{|\lambda|} B_I(\lambda).
\end{aligned} \tag{12.6.31}$$

Since $a_m \to 0$ as $m \to \infty$, we have

$$\tau_m \equiv \sup_{n \geq m} \|\mathcal{M}_{m,n}\| < 1 \tag{12.6.32}$$

for all sufficiently large values of m.

With (12.6.32) and (12.6.27), we have that $u_n^{(\kappa)} \to u_n$ as $\kappa \to \infty$, with a geometric rate of convergence τ_m which is bounded uniformly for $n \geq m$:

$$\left\|u_n - u_n^{(\kappa+1)}\right\|_\infty \leq \tau_m \left\|u_n - u_n^{(\kappa)}\right\|_\infty, \quad \kappa \geq 0. \tag{12.6.33}$$

This completes the proof. □

We can also show from (12.6.27) that the differences of the iterates also converge with this same bound on the geometric rate:

$$\left\|u_n^{(\kappa+1)} - u_n^{(\kappa)}\right\|_\infty \leq \tau_m \left\|u_n^{(\kappa)} - u_n^{(\kappa-1)}\right\|_\infty, \qquad \kappa \geq 1. \tag{12.6.34}$$

From (12.6.33), it is straightforward to show

$$\left\|u_n - u_n^{(\kappa+1)}\right\|_\infty \leq \frac{\tau_m}{1 - \tau_m} \left\|u_n^{(\kappa+1)} - u_n^{(\kappa)}\right\|_\infty, \tag{12.6.35}$$

which can be used to estimate the iteration error, once an estimate or bound for τ_m is known.

12.6.3 The iteration method for the linear system

How do we translate the iteration method (12.6.20)–(12.6.22) for solving $(\lambda - K_n)\, u_n = f$ into an iteration method for the linear system $A_n \boldsymbol{u}_n = \boldsymbol{f}_n$ of (12.6.7)–(12.6.8)? The unknown solution we are seeking is

$$\boldsymbol{u}_n = (u_n(x_{n,1}), \ldots, u_n(x_{n,q_n}))^T .$$

Turning to the iteration formulas (12.6.20)–(12.6.22), assume $\{u_n^{(\kappa)}(x_{n,i})\}$ is known. Begin by calculating the residual $r_n^{(\kappa)}$ at the fine mesh node points:

$$r_n^{(\kappa)}(x_{n,i}) = y(x_{n,i}) - \lambda u_n^{(\kappa)}(x_{n,i}) + \sum_{j=1}^{q_n} w_{n,j} k(x_{n,i}, x_{n,j}) u_n^{(\kappa)}(x_{n,j}) \quad (12.6.36)$$

for $i = 1, \ldots, q_n$. Second, calculate $K_n r_n^{(\kappa)}$ at both the coarse and fine mesh node points:

$$K_n r_n^{(\kappa)}(x) = \sum_{j=1}^{q_n} w_{n,j} k(x, x_{n,j}) r_n^{(\kappa)}(x_{n,j}), \quad x \in \{x_{n,i}\} \cup \{x_{m,i}\}. \quad (12.6.37)$$

Third, calculate the correction $\delta_n^{(\kappa)}$ on the coarse mesh by solving the linear system

$$\lambda \delta_n^{(\kappa)}(x_{m,i}) - \sum_{j=1}^{q_m} w_{m,j} k(x_{m,i}, x_{m,j}) \delta_n^{(\kappa)}(x_{m,j}) = K_n r^{(\kappa)}(x_{m,i}) \quad (12.6.38)$$

for $i = 1, \ldots, q_n$. Fourth, extend this correction to the fine mesh with the Nyström interpolation formula:

$$\delta_n^{(\kappa)}(x_{n,i}) = \frac{1}{\lambda} \left[K_n r_n^{(\kappa)}(x_{n,i}) + \sum_{j=1}^{q_m} w_{m,j} k(x_{n,i}, x_{m,j}) \delta_n^{(\kappa)}(x_{m,j}) \right] \quad (12.6.39)$$

for $i = 1, \ldots, q_n$. Finally, define the new iterate $\boldsymbol{u}_n^{(\kappa+1)}$ on the fine mesh by

$$u_n^{(\kappa+1)}(x_{n,i}) = u_n^{(\kappa)}(x_{n,i}) + \frac{1}{\lambda} \left[r^{(\kappa)}(x_{n,i}) + \delta_n^{(\kappa)}(x_{n,i}) \right], \quad i = 1, \ldots, q_n. \quad (12.6.40)$$

For the iteration of the linear system, we denote

$$\boldsymbol{u}_n^{(\kappa)} = \left(u_n^{(\kappa)}(x_{n,1}), \ldots, u_n^{(\kappa)}(x_{n,q_n}) \right)^T .$$

We stop the iteration when $\left\| \boldsymbol{u}_n - \boldsymbol{u}_n^{(\kappa+1)} \right\|_\infty$ is considered sufficiently small. Generally, we use

$$\nu_\kappa = \frac{\left\| \boldsymbol{u}_n^{(\kappa)} - \boldsymbol{u}_n^{(\kappa-1)} \right\|_\infty}{\left\| \boldsymbol{u}_n^{(\kappa-1)} - \boldsymbol{u}_n^{(\kappa-2)} \right\|_\infty}, \qquad \kappa = 2, 3, \ldots \quad (12.6.41)$$

to estimate the convergence ratio τ_m of (12.6.32). Based on (12.6.33)–(12.6.35), we bound the error by using

$$\left\| \boldsymbol{u}_n - \boldsymbol{u}_n^{(\kappa+1)} \right\|_\infty \leq \frac{\nu_{\kappa+1}}{1-\nu_{\kappa+1}} \left\| \boldsymbol{u}_n^{(\kappa+1)} - \boldsymbol{u}_n^{(\kappa)} \right\|_\infty . \tag{12.6.42}$$

Sometimes in this formula, we replace $\nu_{\kappa+1}$ with the geometric mean of several successive values of ν_κ, to stabilize the ratios when necessary.

Example 12.6.2 Consider solving the equation

$$\lambda u(x) - \int_0^1 k_\gamma(x+y)\, u(y)\, dy = y(x), \qquad 0 \leq x \leq 1 \tag{12.6.43}$$

with

$$k_\gamma(\tau) = \frac{1-\gamma^2}{1+\gamma^2-2\gamma\cos(2\pi\tau)} = 1 + 2\sum_{j=1}^{\infty} \gamma^j \cos(2j\pi\tau)$$

and $0 \leq \gamma < 1$. The eigenvalues and eigenfunctions for the associated integral operator K are

$$\begin{aligned} &\gamma^j, \quad \cos(2j\pi x), \quad j = 0, 1, 2, \ldots, \\ &-\gamma^j, \quad \sin(2j\pi x), \quad j = 1, 2, \ldots . \end{aligned} \tag{12.6.44}$$

This integral equation is obtained when reformulating the Dirichlet problem for Laplace's equation $\Delta u = 0$ on an elliptical region in the plane; see [136, p. 119].

For the integration rule used to define the numerical integration operator K_n, we choose the midpoint rule:

$$\int_0^1 g(y)\, dy \approx h \sum_{j=1}^{n} g((j-1/2)\, h), \quad h = \frac{1}{n}, \ g \in C[0,1]. \tag{12.6.45}$$

For periodic integrands on $[0,1]$, this method converges very rapidly; e.g. see [15, p. 288]. We solve the integral equation (12.6.43) when $\gamma = 0.8$; and we have the unknown functions

$$u_1(x) \equiv 1, \qquad u_2(x) = \sin(2\pi x). \tag{12.6.46}$$

The numerical results are given in Table 12.4, and μ denotes the number of iterates that are computed. In all cases we chose $u_n^{(0)} = 0$. We include a column for the limit of the convergence ratios of (12.6.41),

$$\widetilde{\nu} = \lim_{\kappa \to \infty} \nu_\kappa. \tag{12.6.47}$$

The sequence $\{\nu_\kappa\}$ converged rapidly to an empirical limit $\widetilde{\nu}$, which we give in the table. The ratios $\{\nu_\kappa\}$ are quite well-behaved, as shown empirically by the existence of the limit $\widetilde{\nu}$. □

u_i	λ	m	n	μ	$\|\boldsymbol{u}_n^{(\mu)} - \boldsymbol{u}_n^{(\mu-1)}\|_\infty$	$\|\boldsymbol{u} - \boldsymbol{u}_n^{(\mu)}\|_\infty$	$\widetilde{\nu}$
u_1	−1.00	16	32	10	5.34E−14	7.92E−4	.034
u_1	−1.00	16	64	11	4.57E−14	6.28E−7	.045
u_1	−1.00	16	128	11	5.18E−14	3.96E−13	.045
u_1	−1.00	32	64	6	7.77E−16	6.28E−7	.00080
u_1	−1.00	32	128	6	5.00E−15	3.94E−13	.0013
u_2	−1.00	32	64	7	2.24E−14	6.43E−6	.0053
u_2	−1.00	32	128	7	4.77E−14	4.00E−12	.0060
u_1	0.99	32	64	17	2.00E−14	1.26E−4	.14
u_1	0.99	32	128	17	2.08E−14	7.87E−11	.14

TABLE 12.4. Solving (12.6.43) with iteration (12.6.20)–(12.6.22)

With this two-grid iteration there is a *mesh independence principle* working. This means that as n increases, the number of iterates to be computed in order to obtain a given reduction in the initial error is essentially independent of n. This is in contrast to what usually occurs with most iteration methods for solving finite difference discretizations of partial differential equations, where the number of iterates to be computed increases as the parametrization variable n increases. The present *mesh independence* can be supported theoretically. The convergence operator $\mathcal{M}_{m,n}$ of (12.6.28) can be shown to satisfy

$$\lim_{n\to\infty} \mathcal{M}_{m,n} = \frac{1}{\lambda}(\lambda - K_m)^{-1}(K - K_m)K. \tag{12.6.48}$$

The norm of the limiting value is less than 1 if m is chosen sufficiently large; and thus the number of iterates to be computed can be shown to be independent of n, for n chosen sufficiently large. Also, the rate of convergence is improved by increasing m.

The linear systems solved by iteration in the preceding examples were relatively small, and they could have been solved more simply by a direct method such as Gaussian elimination. In contrast, consider the radiosity equation

$$\lambda\rho(P) - \int_S \rho(Q)\frac{\partial}{\partial\nu_Q}\left(|P-Q|^2\right) dS_Q = \psi(P), \quad P \in S$$

which arises in computer graphics. The region S is usually a complicated polyhedral surface in space. Discretization of this equation ordinarily leads to linear systems of order $q_n \geq 10,000$; and iteration methods are the only practical means of solving the equation.

12.6.4 An operations count

We will look at the number of arithmetic operations used in computing a single iteration of (12.6.36)–(12.6.40). In doing so, we assume such quanti-

ties as $\{f(x_{n,i})\}$ and $\{w_{n,i}k(x_{n,i}, x_{n,j})\}$ have already been calculated and saved for later use in the iteration.

1. To calculate the residuals $\{r_n^{(\kappa)}(x_{n,i})\}$ of (12.6.36) requires approximately $2q_n^2$ arithmetic operations (combining additions, subtractions, multiplications, and divisions).

2. To evaluate $\{K_n r_n^{(\kappa)}(x_{n,i})\}$ and $\{K_n r_n^{(\kappa)}(x_{m,i})\}$ requires approximately $2q_n(q_n + q_m)$ arithmetic operations.

3. To solve the linear system (12.6.38) for $\{\delta_n^{(\kappa)}(x_{m,i})\}$ requires approximately $2q_m^2$ arithmetic operations, provided an LU-factorization of the matrix A_m has already been calculated and saved.

4. To evaluate $\{\delta_n^{(\kappa)}(x_{n,i})\}$ using the Nyström interpolation formula (12.6.39) requires approximately $2q_n q_m$ arithmetic operations.

5. The final step (12.6.40) requires only $3q_n$ arithmetic operations, and is negligible in comparison to the other costs.

Combining these, we have a total cost of approximately

$$2q_n^2 + 2(q_n + q_m)^2 \tag{12.6.49}$$

arithmetic operations. For $q_n \gg q_m$, this iteration method has a cost of approximately $4q_n^2$ arithmetic operations per iteration. This is quite reasonable in most situations.

The subject of iteration methods for solving integral equations is a large one. For a more complete introduction, see [18, Chapter 6]. Additional results for two-grid iteration, including collocation methods, are given in [17, 103]. Multigrid methods for integral equations are introduced in [104], and so-called 'fast methods' are given in [105, 106].

Exercise 12.6.1 Recall (12.6.17) and assume that e_n is restricted to lay in a bounded set about the origin. Show that δ_n is then restricted to lay in a compact set about the origin.

Exercise 12.6.2 Consider the iteration method (12.6.12)–(12.6.13). Derive for it the convergence formula analogue of (12.6.27)–(12.6.28)

$$u_n - u_n^{(\kappa+1)} = \mathcal{M}_{m,n}\left[u_n - u_n^{(\kappa)}\right].$$

Exercise 12.6.3 Continuing with Exercise 12.6.2, prove that

$$\sup_{n>m} \left\|\mathcal{M}_{m,n}^2\right\| \to 0 \quad \text{as } n \to \infty.$$

Exercise 12.6.4 Carry out the analogue of (12.6.36)–(12.6.40) for the iteration method (12.6.14)–(12.6.15).

Exercise 12.6.5 Continuing with Exercise 12.6.4, do an operations count of the iteration method (12.6.19), in analogy with obtaining the operations count in (12.6.14)–(12.6.15).

Exercise 12.6.6 Extend the methods (12.6.14)–(12.6.15) and (12.6.20)–(12.6.22) to the product integration deiscretization schemes of Section 12.5.

12.7 Projection methods for nonlinear equations

Recall the material of Sections 5.3–5.5 of Chapter 5 on nonlinear fixed point problems. We will define and analyze projection methods for the discretization of fixed point problems

$$u = T(u) \tag{12.7.1}$$

with $T : H \subset V \to V$ a completely continuous nonlinear operator. The space V is a Banach space, and H is an open subset of V. The prototype example of T is the Urysohn integral equation of Example 5.3.10:

$$T(u)(t) = g(t) + \int_a^b k(t, s, u(s))\, ds. \tag{12.7.2}$$

The function $k(t, s, u)$ is to possess such properties as to ensure it is a completely continuous operator on some open set $H \subset C[a, b]$ (see Section 5.3.10).

Recall the theoretical framework of Subsection 12.1.3. We define the projection method for solving (12.7.1) as follows. For a given discretization parameter n, find $u_n \in V_n$ satisfying the equation

$$u_n = P_n T(u_n). \tag{12.7.3}$$

We can illustrate the method in analogy with Section 12.2, but defer this to later in this section.

There are two major approaches to the error analysis of (12.7.3): (1) Linearize the problem and apply Theorem 5.1.3, the Banach fixed point theorem; (2) Apply the theory associated with the rotation of a completely continuous vector field (see Section 5.5).

12.7.1 Linearization

We begin the linearization process by discussing the error in the linearization of $T(v)$ about a point v_0:

$$R(v; v_0) \equiv T(v) - [T(v_0) + T'(v_0)(v - v_0)] \tag{12.7.4}$$

Lemma 12.7.1 *Let V be a Banach space, and let H be an open subset of V. Let $T : H \subset V \to V$ be twice continuously differentiable with $T''(v)$ bounded over any bounded subset of H. Let $B \subset H$ be a closed, bounded, and convex set with a non-empty interior. Let v_0 belong to the interior of B, and define $R(v; v_0)$, as above. Then for all $v_1, v_2 \in B$,*

$$\|R(v_2; v_1)\| \leq \frac{1}{2} M \|v_1 - v_2\|^2 \tag{12.7.5}$$

with $M = \sup_{v \in B} \|T''(v)\|$. Moreover,

$$\|T'(v_2) - T'(v_1)\| \leq M \|v_2 - v_1\|, \tag{12.7.6}$$

implying $T'(v)$ is Lipschitz continuous; and

$$\|R(v_1; v_0) - R(v_2; v_0)\| \leq M \left[\|v_1 - v_0\| + \tfrac{1}{2} \|v_1 - v_2\|\right] \|v_1 - v_2\|. \tag{12.7.7}$$

Proof. The result (12.7.5) is immediate from Proposition 5.3.13 of Section 5.3; and the proof of (12.7.6) can be based on Proposition 5.3.11 when applied to $T'(v)$. The proof of (12.7.7) is let as an exercise. □

As earlier, assume $T : H \subset V \to V$ is a completely continuous nonlinear operator. Assume (12.7.1) has an isolated solution $u^* \in H$, and assume it is unique within the ball

$$B(u^*, \epsilon) = \{v \mid \|v - u^*\| \leq \epsilon\}$$

for some $\epsilon > 0$ and with $B(u^*, \epsilon) \subset H$. We assume T is twice continuously differentiable over H, with $T''(v)$ uniformly bounded over all bounded neighborhoods, such as $B(u^*, \epsilon)$:

$$M(u^*, \epsilon) \equiv \sup_{v \in B(u^*, \epsilon)} \|T''(v)\| < \infty.$$

Assume that 1 is not an eigenvalue of $T'(u^*)$. This then implies that $I - T'(u^*)$ is a bijective mapping from V to V and that it has a bounded inverse. For a proof, invoke Proposition 5.5.5 to show $T'(u^*)$ is a compact linear operator, and then apply Theorem 2.8.10, the Fredholm alternative theorem. Henceforth, we let $L = T'(u^*)$.

Assume that the projections $\{P_n\}$ are pointwise convergent to the identity on V,

$$P_n v \to v \quad \text{as } n \to \infty, \quad \forall\, v \in V. \tag{12.7.8}$$

Then from Proposition 5.5.5 and Lemma 12.1.4,

$$\|(I - P_n) L\| \to 0 \quad \text{as } n \to \infty.$$

From Theorem 12.1.2, $(I - P_n L)^{-1}$ exists for all sufficiently large n and is uniformly bounded with respect to all such n.

We want to show that for all sufficiently large n, (12.7.3) has a unique solution within $B(u^*, \epsilon_1)$ for some $0 < \epsilon_1 \leq \epsilon$. We also would like to obtain bounds on the rate of convergence of u_n to u^*. In (12.7.3), expand $T(u_n)$ about u^*, obtaining

$$T(u_n) = T(u^*) + L\,(u_n - u^*) + R(u_n; u^*).$$

Equation (12.7.3) can be rewritten as the equivalent equation

$$(I - P_n L)\,(u_n - u^*) = P_n u^* - u^* + P_n R(u_n; u^*) \tag{12.7.9}$$

Introduce a new unknown $\delta_n = u_n - u^*$, and then write

$$\begin{aligned} \delta_n &= (I - P_n L)^{-1}\,(P_n u^* - u^*) + (I - P_n L)^{-1}\,R(\delta_n + u^*; u^*) \\ &\equiv F_n(\delta_n). \end{aligned} \tag{12.7.10}$$

We are interested in showing that on some ball about the origin in V, of radius $\epsilon_1 \leq \epsilon$, this fixed point equation has a unique solution δ_n, provided only that n is chosen sufficiently large. This can be done by showing that F_n is a contractive mapping on a ball $B(0, \epsilon_1)$ provided that $\epsilon_1 > 0$ is chosen sufficiently small. To do this requires showing the two main hypotheses of Theorem 5.1.3, the Banach contractive mapping theorem. Namely, show that if n is sufficiently large, there exists ϵ_1 for which:

1.
$$F_n : B(0, \epsilon_1) \to B(0, \epsilon_1); \tag{12.7.11}$$

2.
$$\|F_n(\delta_{n,1}) - F_n(\delta_{n,2})\| \leq \alpha\,\|\delta_{n,1} - \delta_{n,2}\|\,, \quad \delta_{n,1}, \delta_{n,2} \in B(0, \epsilon_1) \tag{12.7.12}$$
with $\alpha < 1$ and independent of n, provided n is chosen to be sufficiently large.

The number ϵ_1 can be made independent of n, provided n is sufficiently large. These two properties can be proven using the various results and assumptions we have made regarding T and $\{P_n\}$, and we leave their demonstration as an exercise for the reader. This proves that for all sufficiently large n, the approximating equation (12.7.3) has a unique solution u_n in some ball of fixed radius about u^*.

There are a number of results on the rate of convergence of u_n to u^*, and we quote only one of them. With the same hypotheses on T and $\{P_n\}$ as above,

$$\|u^* - u_n\|_V \leq \|\,[I - T'(u^*)]^{-1}\,\|\,(1 + \gamma_n)\,\|u^* - P_n u^*\|_V \tag{12.7.13}$$

with $\gamma_n \to 0$ as $n \to \infty$. A proof of this result is given in [25, Theorem 2.2]. This error bound is somewhat comparable to the bound (12.1.24) given earlier for linear projection methods; also see Exercise 12.1.3.

12.7.2 A homotopy argument

This mode of analysis of projection methods for the discretization of fixed point problems (12.7.1) requires fewer assumptions on the nonlinear operator T, and there is no assumption on the differentiability of T. As before, we assume $T : H \subset V \to V$ is a completely continuous operator. Let u^* be an isolated fixed point of T, and assume u^* is isolated within the ball $B(u^*, \epsilon)$ for some $\epsilon > 0$. Further, assume that u^* has a nonzero index (recall the discussion of *index* as discussed in **P3** of Subsection 5.5.1 in Chapter 5). The discussion in **P4** of Subsection 5.5.1 assures us that the index of u^* is nonzero if $I - T'(u^*)$ is a bijective linear operator; but the index can be nonzero under weaker assumptions on u^*; for example, see **P5** of Subsection 5.5.1.

Let S denote the boundary of $B(u^*, \epsilon)$. Recalling Subsection 5.5.1, we have the concept of the quantity $\text{Rot}(\Phi)$, the *rotation* of the completely continuous vector field

$$\Phi(v) = v - T(v), \quad v \in B(u^*, \epsilon).$$

Also, introduce the approximating vector field

$$\Phi_n(v) = v - P_n T(v), \quad v \in B(u^*, \epsilon).$$

By our assumptions on u^*, $\Phi(v) \neq 0$ for all $v \in S$, and consequently $\text{Rot}(\Phi) \neq 0$ (and in fact equals the index of the fixed point u^*). We introduce the homotopy

$$X(v,t) = v - (1-t)T(v) - tP_n T(v), \quad v \in B(u^*, \epsilon) \tag{12.7.14}$$

for $0 \leq t \leq 1$. We show that for all sufficiently large values of n, say $n \geq N(\epsilon)$, this homotopy satisfies the hypotheses of **P2** of Subsection 5.5.1; and consequently, the index of Φ_n will be the same as that of Φ, namely nonzero. In turn, this implies that Φ_n contains zeros within the ball $B(u^*, \epsilon)$, or equivalently, the approximating equation (12.7.3) has solutions within this ϵ-neighborhood of u^*.

Recalling the four hypotheses of **P2** of Subsection 5.5.1, only the fourth one is difficult to show, namely that

$$X(v,t) \neq 0, \quad \forall\, v \in S,\ 0 \leq t \leq 1 \tag{12.7.15}$$

for all sufficiently large values of n. To examine this, rewrite (12.7.14) as

$$X(v,t) = [v - T(v)] + t\,[T(v) - P_n T(v)]\,. \tag{12.7.16}$$

We note as a preliminary lemma that

$$\alpha \equiv \inf_{v \in S} \|v - T(v)\| > 0. \tag{12.7.17}$$

To prove this, assume the contrary. Then there exists a sequence $\{v_m\} \subset S$ for which

$$v_m - T(v_m) \to 0 \quad \text{as } m \to \infty. \tag{12.7.18}$$

Since S is bounded and T is completely continuous, the sequence $\{T(v_m)\}$ has a convergent subsequence, say

$$T(v_{m_j}) \to w \quad \text{as } m_j \to \infty.$$

When combined with (12.7.18), this implies $v_{m_j} \to w$; and the closedness of S then implies $w \in S$. The continuity of T implies $v = T(v)$, contradicting the assumption that S contains no fixed points of S. This proves (12.7.17).

Returning to (12.7.16), we have

$$\|X(v,t)\| \geq \alpha - t\,\|T(v) - P_n T(v)\|\,, \quad v \in S,\ 0 \leq t \leq 1. \tag{12.7.19}$$

We assert that

$$\sup_{v \in S} \|T(v) - P_n T(v)\| \to 0 \quad \text{as } n \to \infty.$$

This follows by writing this in the equivalent form

$$\sup_{w \in T(S)} \|w - P_n w\| \to 0 \quad \text{as } n \to \infty.$$

This results follows from Lemma 12.1.3, (12.7.8), and the precompactness of $T(S)$.

When combined with (12.7.19), we have that for all sufficiently large n, say $n \geq N(\epsilon)$,

$$\|X(v,t)\| \geq \frac{\alpha}{2}, \quad v \in S,\ 0 \leq t \leq 1.$$

This completes the proof of (12.7.15), the fourth hypothesis of **P2** of Subsection 5.5.1. As discussed earlier, this implies that (12.7.3) has solutions within $B(u^*, \epsilon)$. As we make $\epsilon \to 0$, this construction also implies the existence of a sequence of approximating solutions u_n which converges to u^* as $n \to \infty$. The analysis of the preceding few paragraphs is essentially the argument given in Krasnoselskii [147, Section 3.3] for the convergence of Galerkin's method for solving (12.7.1).

This is a powerful means of argument for the existence and convergence of approximation solutions of a completely continuous fixed point problem. But it does not imply that the equations (12.7.3) are uniquely solvable, and indeed they may not be. For an example in which $v = T(v)$ has a isolated fixed point u^*, but one for which the approximating equations are not *uniquely* solvable in any neighborhood of u^*, see [14, p. 590].

12.7.3 The approximating finite-dimensional problem

Consider solving the Urysohn nonlinear equation

$$u(x) = f(x) + \int_D k(x, y, u(y))\, dy \equiv T(u)(x), \quad x \in D$$

for an integration region $D \subset \mathbb{R}^d$. We denote by V the function space for the consideration of this equation, and we let V_n denote the finite-dimensional subspace from which our approximation will be chosen,

$$u_n(x) = \sum_{j=1}^{\kappa_n} c_j \phi_j(x), \quad x \in D. \tag{12.7.20}$$

In this discussion, recall the general framework of Section 12.1 and the specific examples of Section 12.2.

To be more specific, let V be a space of continuous functions, and let P_n be an interpolatory projection operator from V to V_n, based on node points $\{x_j \mid 1 \le j \le \kappa_n\}$. Then the approximating equation (12.7.3) is equivalent to choosing u_n as in (12.7.20) with $\{c_j\}$ satisfying the nonlinear system

$$\sum_{j=1}^{\kappa_n} c_j \phi_j(x_i) = f(x_i) + \int_D k\Big(x_i, y, \sum_{j=1}^{\kappa_n} c_j \phi_j(y)\Big)\, dy, \quad i = 1, \ldots, \kappa_n. \tag{12.7.21}$$

This is a nontrivial system to solve, and usually some variant of Newton's method is used to find an approximating solution. From a practical perspective, a major difficulty is that the integral will need to be numerically evaluated repeatedly with varying x_i and varying iterates $\{c_j^{(k)}\}$, where k is an index for the iterative solution of the system.

An important variant is possible for the Hammerstein equation

$$u(x) = f(x) + \int_D k(x, y) g(y, u(y))\, dy \equiv T(u)(x), \quad x \in D. \tag{12.7.22}$$

We can convert this problem to one for which the number of needed numerical integrations is reduced greatly. Introduce a new unknown

$$w(x) = g(x, u(x)), \quad x \in D.$$

Then u can be recovered from w using

$$u(x) = f(x) + \int_D k(x, y) w(y)\, dy, \quad x \in D. \tag{12.7.23}$$

To solve for w, use the equation

$$w(x) = g\Big(x, f(x) + \int_D k(x, y) w(y)\, dy\Big), \quad x \in D. \tag{12.7.24}$$

If we examine the nonlinear system (12.7.21) for this equation, we can greatly minimize the needed integrations, needing only the evaluations of the integrals

$$\int_D k(x_i, y)\phi_j(y)\, dy, \quad i, j = 1, \ldots, \kappa_n.$$

The integrations need not be recomputed for each iteration of the solution of the nonlinear system. We leave the further analysis of this to Exercise 12.7.4. For further literature on this method, see [152].

Exercise 12.7.1 Prove (12.7.7) of Lemma 12.7.1.
Hint: Use the definition (12.7.4) to write out both $R(v_1; v_0)$ and $R(v_2; v_0)$. Simplify; then apply (12.7.5) and (12.7.6).

Exercise 12.7.2 Prove (12.7.11) and (12.7.12), provided that $\epsilon_1 > 0$ is chosen sufficiently small and n is chosen sufficiently large.

Exercise 12.7.3 Using (12.7.9), prove a weaker form of (12.7.13), namely

$$\|u^* - u_n\|_V \le c \|u^* - P_n u^*\|_V, \quad n \ge N$$

for some $N \ge 1$, with c a constant (dependent on N).

Exercise 12.7.4 Fill in the details of the solution of the nonlinear system (12.7.21) for the equation (12.7.24).

Exercise 12.7.5 Do a detailed presentation and analysis of the solution of

$$u(t) = g(t) + \int_a^b k(t, s, u(s))\, ds, \quad a \le t \le b$$

using piecewise linear collocation (as in Subsection 12.2.1). Include a discussion of the nonlinear system which you must setup and solve.

Exercise 12.7.6 Repeat Exercise 12.7.5 for the equations (12.7.23)–(12.7.24).

Exercise 12.7.7 Recall the material of Section 12.3 on iterated projection methods. Define the iterated projection solution for 12.7.3 as

$$\widehat{u}_n = T(u_n).$$

Show $P_n \widehat{u}_n = u_n$ and $\widehat{u}_n = T(P_n \widehat{u}_n)$. This can be used as a basis for a direct analysis of the convergence of $\{\widehat{u}_n\}$.

Suggestion for Further Reading.

Parts of this chapter are a modification of portions of the presentation in Atkinson [18, Chaps. 3, 4]. Another introduction to the numerical solution of integral equations is given in Kress [149]. The first general treatment of projection methods appears to have been due to L.V. Kantorovich in

1948, and those arguments appear in an updated form in KANTOROVICH AND AKILOV [135]. The general theory of collectively compact operator approximations was created by P. ANSELONE, and the best introduction to it is his book [5]. For a survey of numerical methods for solving nonlinear integral equations, see ATKINSON [16]. Extensions of the ideas of Section 12.7 to Nyström's method for nonlinear equations are given in ATKINSON [12] and ATKINSON AND POTRA [25].

13
Boundary Integral Equations

In Chapter 10, we examined finite element methods for the numerical solution of Laplace's equation. In this chapter, we propose an alternative approach. We introduce the idea of reformulating Laplace's equation as a *boundary integral equation (BIE)*, and then we consider the numerical solution of Laplace's equation by numerically solving its reformulation as a BIE. Some of the most important boundary value problems for elliptic partial differential equations have been studied and solved numerically by this means; and depending on the requirements of the problem, the use of BIE reformulations may be the most efficient means of solving these problems. Examples of other equations solved by use of BIE reformulations are the Helmholtz equation ($\Delta u + \lambda u = 0$) and the biharmonic equation ($\Delta^2 u = 0$). We consider here the use of boundary integral equations in solving only planar problems for Laplace's equation. For the domain D for the equation, we restrict it or its complement to be a simply-connected set with a smooth boundary S. Most of the results and methods given here will generalize to other equations (e.g. Helmholtz's equation).

In this chapter, Section 13.1 contains a theoretical framework for BIE reformulations of Laplace's equation in $\mathbb{R}^2$, giving the most popular of such boundary integral equations. For much of the history of BIE, those of the second kind have been the most popular; this includes the work of Ivar Fredholm, Carl Neumann, David Hilbert, and others in the late 1800s and early 1900s. In Section 13.2, we discuss the numerical solution of such BIE of the second kind. In Section 13.3, we introduce briefly the study of BIE of the first kind, and we discuss the use of Fourier series as a means of studying these equations and numerical methods for their solution.

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_13,
© Springer Science + Business Media, LLC 2009

As in the preceding Chapter 12, here we use notation that is popular in the literature on boundary integral equations.

13.1 Boundary integral equations

Let D be a bounded open simply-connected region in the plane, and let its boundary be denoted by S. At a point $P \in S$, let $\boldsymbol{n}_P$ denote the inner unit normal to S. We restate the principal boundary value problems of interest when solving Laplace's equation on D.

The Interior Dirichlet Problem: Find $u \in C(\overline{D}) \cap C^2(D)$ that satisfies

$$\begin{aligned} \Delta u(P) &= 0, \qquad P \in D \\ u(P) &= f(P), \qquad P \in S \end{aligned} \tag{13.1.1}$$

with $f \in C(S)$ a given boundary function.

The Interior Neumann Problem: Find $u \in C^1(\overline{D}) \cap C^2(D)$ that satisfies

$$\begin{aligned} \Delta u(P) &= 0, \qquad P \in D \\ \frac{\partial u(P)}{\partial \boldsymbol{n}_P} &= f(P), \qquad P \in S \end{aligned} \tag{13.1.2}$$

with $f \in C(S)$ a given boundary function.

Another important boundary value problem is that with a mixture of Neumann and Dirichlet boundary conditions on different sections of the boundary, or perhaps some combination of them. The techniques introduced here can also be used to study and solve such mixed boundary value problems, but we omit any such discussion here. Corresponding to the interior Dirichlet and Neumann problems given above, there are corresponding exterior problems. These are discussed later in the section. Functions satisfying Laplace's equation are often called "harmonic functions". The study of Laplace's equation is often referred to as "potential theory", since many applications involve finding a potential function u in order to construct a conservative vector field ∇u.

The above boundary value problems have been discussed earlier, in Chapter 8. Here we give a theorem summarizing the main results on their solvability, in the form needed here.

Theorem 13.1.1 *Let the function $f \in C(S)$; and assume S can be parameterized by a twice continuously differentiable function. Then:*

1. *The Dirichlet problem* (13.1.1) *has a unique solution.*

2. *The Neumann problem* (13.1.2) *has a unique solution, up to the addition of an arbitrary constant, provided*

$$\int_S f(Q)\,dS = 0. \tag{13.1.3}$$

13.1.1 Green's identities and representation formula

A very important tool for studying elliptic partial differential equations is the *divergence theorem* or *Gauss's theorem*. This was given earlier in Section 7.6 of Chapter 7 (Proposition 7.6.1); but we re-state it in the form needed for the planar Laplace equation, a form usually called "Green's Theorem". We state the result for regions Ω that are not simply-connected and whose boundaries need not be smooth. This form is needed when proving Green's representation formula (13.1.23).

Let Ω denote an open planar region. Let its boundary Γ consist of $m+1$ distinct simple closed curves, $m \geq 0$,

$$\Gamma = \Gamma_0 \cup \cdots \cup \Gamma_m.$$

Assume $\Gamma_1, \ldots, \Gamma_m$ are contained in the interior of Γ_0. For each $i = 1, \ldots, m$, let Γ_i be exterior to the remaining curves $\Gamma_1, \ldots, \Gamma_{i-1}, \Gamma_{i+1}, \ldots, \Gamma_m$. Further, assume each curve Γ_i is a piecewise smooth curve. We say a curve γ is *piecewise smooth* if:

1. It can be broken into a finite set of curves $\gamma_1, \ldots, \gamma_k$ with each γ_j having a parametrization which is at least twice continuously differentiable.

2. The curve γ does not contain any cusps, meaning that each pair of adjacent curves γ_i and γ_{i+1} join at an interior angle in the interval $(0, 2\pi)$.

The region Ω is interior to Γ_0, but it is exterior to each of the curves $\Gamma_1, \ldots, \Gamma_m$. The orientation of Γ_0 is to be counterclockwise, while the curves $\Gamma_1, \ldots, \Gamma_m$ are to be clockwise.

Theorem 13.1.2 (The divergence theorem) *Assume* $\boldsymbol{F} : \overline{\Omega} \to \mathbb{R}^2$ *with each component of* $\boldsymbol{F}$ *contained in* $C^1(\overline{\Omega})$. *Then*

$$\int_\Omega \nabla \cdot \boldsymbol{F}(Q)\,d\Omega = -\int_\Gamma \boldsymbol{F}(Q) \cdot \boldsymbol{n}(Q)\,d\Gamma. \tag{13.1.4}$$

This important result, which generalizes the fundamental theorem of the calculus, is proven in most standard textbooks on "advanced calculus". It is also a special case of Proposition 7.6.1 from Chapter 7.

Using the divergence theorem, one can obtain *Green's identities* and *Green's representation formula*. Assuming $u \in C^1(\overline{\Omega})$ and $w \in C^2(\overline{\Omega})$, one can prove *Green's first identity* by letting $\boldsymbol{F} = u\nabla w$ in (13.1.4):

$$\int_\Omega u\Delta w \, d\Omega + \int_\Omega \nabla u \cdot \nabla w \, d\Omega = -\int_\Gamma u \frac{\partial w}{\partial \boldsymbol{n}} \, d\Gamma. \tag{13.1.5}$$

This was given earlier in (7.6.4) of Chapter 7.

Next, assume $u, w \in C^2(\overline{\Omega})$. Interchanging the roles of u and w in (13.1.5), and then subtracting the two identities, one obtains *Green's second identity*:

$$\int_\Omega (u\Delta w - w\Delta u) \, d\Omega = \int_\Gamma \left(w\frac{\partial u}{\partial \boldsymbol{n}} - u\frac{\partial w}{\partial \boldsymbol{n}} \right) d\Gamma. \tag{13.1.6}$$

The identity (13.1.5) can be used to prove (i) if the Neumann problem (13.1.2) has a solution, then it is unique up to the addition of an arbitrary constant; and (ii) if the Neumann problem is to have a solution, then the condition (13.1.3) is necessary. The identity (13.1.5) also leads to a proof of the uniqueness of possible solutions of the Dirichlet problem.

Return to the original domain D on which the problems (13.1.1) and (13.1.2) are posed, and assume $u \in C^2(\overline{D})$. Let $u(Q)$ be a solution of Laplace's equation, and let $w(Q) = \log|A - Q|$, with $A \in D$. Here $|A - Q|$ denotes the ordinary Euclidean length of the vector $A - Q$. Define Ω to be D after removing the small disk $B(A, \epsilon) \equiv \{Q \mid |A - Q| \le \epsilon\}$, with $\epsilon > 0$ so chosen that $B(A, 2\epsilon) \subset D$. Note that for the boundary Γ of Ω,

$$\Gamma = S \cup \{Q \mid |A - Q| = \epsilon\}.$$

Apply (13.1.6) with this choice of Ω, and then let $\epsilon \to 0$. Doing so, and then carefully computing the various limits, we obtain *Green's representation formula*:

$$u(A) = \frac{1}{2\pi} \int_S \left[\frac{\partial u(Q)}{\partial \boldsymbol{n}_Q} \log|A - Q| - u(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|A - Q| \right] dS_Q, \quad A \in D. \tag{13.1.7}$$

This expresses u over D in terms of the boundary values of u and its normal derivative on S.

From hereon in this chapter, we assume S has a parametrization $\boldsymbol{r}(t)$ which is in C^2. Some of the results given here are still true if S is only piecewise smooth; but we refer to [18, Chapters 7–9] for a more complete treatment.

We can take limits in (13.1.7) as A approaches a point on the boundary S. Let $P \in S$. Then after a careful calculation,

$$\lim_{A\to P} \int_S \frac{\partial u(Q)}{\partial \boldsymbol{n}_Q} \log|A-Q|\, dS_Q = \int_S \frac{\partial u(Q)}{\partial \boldsymbol{n}_Q} \log|P-Q|\, dS_Q,$$
$$\lim_{\substack{A\to P \\ A\in D}} \int_S u(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|A-Q|\, dS_Q$$
$$= -\pi u(P) + \int_S u(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|P-Q|\, dS_Q. \tag{13.1.8}$$

A proof of (13.1.8), and of the associated limit in (13.1.26), can be found in [57, pp. 197–202] or in many other texts on Laplace's equation.

Using these limits in (13.1.7) yields the relation

$$u(P) = \frac{1}{\pi} \int_S \left[\frac{\partial u(Q)}{\partial \boldsymbol{n}_Q} \log|P-Q| - u(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|P-Q| \right] dS_Q, \quad P \in S \tag{13.1.9}$$

which gives a relationship between the values of u and its normal derivative on S.

The formula (13.1.9) is an example of a boundary integral equation; and it can be used to create other such boundary integral equations. First, however, we need to look at solving Laplace's equation on exterior regions $D_e = \mathbb{R}^2 \backslash \overline{D}$ and to obtain formulas that correspond to (13.1.7)–(13.1.9) for such exterior regions. We also use the notation $D_i = D$ in some places, to indicate clearly that an interior region is being used.

13.1.2 The Kelvin transformation and exterior problems

Define a transformation $\mathcal{T} : \mathbb{R}^2 \backslash \{\mathbf{0}\} \to \mathbb{R}^2 \backslash \{\mathbf{0}\}$,

$$\mathcal{T}(x,y) = (\xi,\eta) \equiv \frac{1}{r^2}(x,y), \qquad r = \sqrt{x^2+y^2}. \tag{13.1.10}$$

In polar coordinates,

$$\mathcal{T}(r\cos\theta, r\sin\theta) = \frac{1}{r}(\cos\theta, \sin\theta).$$

Thus a point (x,y) is mapped onto another point (ξ,η) on the same ray emanating from the origin, and we call (ξ,η) the inverse of (x,y) with respect to the unit circle. Note that $\mathcal{T}(\mathcal{T}(x,y)) = (x,y)$, so that $\mathcal{T}^{-1} = \mathcal{T}$. The Jacobian matrix for $\mathcal{T}$ is

$$J(\mathcal{T}) = \begin{pmatrix} \dfrac{\partial \xi}{\partial x} & \dfrac{\partial \xi}{\partial y} \\ \dfrac{\partial \eta}{\partial x} & \dfrac{\partial \eta}{\partial y} \end{pmatrix} = \frac{1}{r^2} H \tag{13.1.11}$$

with

$$H = \begin{pmatrix} \dfrac{y^2 - x^2}{r^2} & \dfrac{-2xy}{r^2} \\ \dfrac{-2xy}{r^2} & \dfrac{x^2 - y^2}{r^2} \end{pmatrix}.$$

The matrix H is orthogonal with determinant (-1), and

$$\det J(\mathcal{T}(x,y)) = -\frac{1}{r^2}.$$

Assume the bounded open region $D \equiv D_i$ contains the origin $\mathbf{0}$. For a function $u \in C(\overline{D}_e)$, define

$$\widehat{u}(\xi,\eta) = u(x,y), \quad (\xi,\eta) = \mathcal{T}(x,y),\ (x,y) \in \overline{D}_e. \tag{13.1.12}$$

This is called the *Kelvin transformation* of u. Introduce the interior region $\widehat{D} = \mathcal{T}(D_e)$, and let $\widehat{S}$ denote the boundary of $\widehat{D}$. The boundaries S and $\widehat{S}$ have the same degree of smoothness. In addition, the condition $(\xi,\eta) \to \mathbf{0}$ in $\widehat{D}$ corresponds to $r \to \infty$ for points $(x,y) \in D_e$. For a function u satisfying Laplace's equation on D, it is a straightforward calculation to show

$$\Delta\widehat{u}(\xi,\eta) = r^4 \Delta u(x,y) = 0, \quad (\xi,\eta) = \mathcal{T}(x,y),\ (x,y) \in D_e \tag{13.1.13}$$

thus showing $\widehat{u}$ to be harmonic on $\widehat{D}$. We can pass from the solution of Laplace's equation on the unbounded region D_e to the bounded open region $\widehat{D}$.

If we were to impose the Dirichlet condition $u = f$ on the boundary S, this is equivalent to the Dirichlet condition

$$\widehat{u}(\xi,\eta) = f(\mathcal{T}^{-1}(\xi,\eta)), \quad (\xi,\eta) \in \widehat{S}.$$

From the existence and uniqueness result of Theorem 13.1.1, the interior Dirichlet problem on $\widehat{D}$ will have a unique solution. This leads us to considering the following problem.

The Exterior Dirichlet Problem: Find $u \in C(\overline{D}_e) \cap C^2(D_e)$ that satisfies

$$\begin{aligned} \Delta u(P) &= 0, & P &\in D_e, \\ u(P) &= f(P), & P &\in S, \\ \lim_{r\to\infty} \sup_{|P|\geq r} |u(P)| &< \infty, \end{aligned} \tag{13.1.14}$$

with $f \in C(S)$ a given boundary function.

Using the above discussion on the Kelvin transform, this converts to the interior Dirichlet problem

$$\begin{aligned} \Delta\widehat{u}(\xi,\eta) &= 0, & (\xi,\eta) &\in \widehat{D} \\ u(\xi,\eta) &= f(\mathcal{T}^{-1}(\xi,\eta)), & (\xi,\eta) &\in \widehat{S} \end{aligned} \tag{13.1.15}$$

and Theorem 13.1.1 guarantees the unique solvability of this problem. The condition on $u(x,y)$ as $r \to \infty$ can be used to show that $\widehat{u}(\xi,\eta)$ has a removable singularity at the origin; and $\widehat{u}(0,0)$ will be the value of $u(x,y)$ as $r \to \infty$. Thus the above exterior Dirichlet problem has a unique solution.

For functions $u \in C^1(\overline{D}_e)$,

$$\frac{\partial u(x,y)}{\partial \boldsymbol{n}(x,y)} = -\rho^2 \frac{\partial \widehat{u}(\xi,\eta)}{\partial \widehat{\boldsymbol{n}}(\xi,\eta)}, \quad \rho = \frac{1}{r} = \sqrt{\xi^2+\eta^2} \tag{13.1.16}$$

with $\widehat{\boldsymbol{n}}(\xi,\eta)$ the unit interior normal to $\widehat{S}$ at (ξ,η). Thus the Neumann condition

$$\frac{\partial u(x,y)}{\partial \boldsymbol{n}(x,y)} = f(x,y), \quad (x,y) \in S$$

is equivalent to

$$\frac{\partial \widehat{u}(\xi,\eta)}{\partial \widehat{\boldsymbol{n}}(\xi,\eta)} = -\frac{1}{\rho^2} f(\mathcal{T}^{-1}(\xi,\eta)) \equiv \widehat{f}(\xi,\eta), \quad (\xi,\eta) \in \widehat{S}. \tag{13.1.17}$$

Also,

$$\int_S \frac{\partial u}{\partial \boldsymbol{n}} dS = -\int_{\widehat{S}} \frac{\partial \widehat{u}}{\partial \widehat{\boldsymbol{n}}} d\widehat{S}. \tag{13.1.18}$$

Using this information, consider the following problem.

The Exterior Neumann Problem. Find $u \in C^1(\overline{D}_e) \cap C^2(D_e)$ that satisfies

$$\begin{aligned} \Delta u(P) &= 0, && P \in D_e \\ \frac{\partial u(P)}{\partial \boldsymbol{n}_P} &= f(P), && P \in S \end{aligned} \tag{13.1.19}$$

$$u(r\cos\theta, r\sin\theta) = \mathcal{O}\left(\frac{1}{r}\right), \quad \frac{\partial u(r\cos\theta, r\sin\theta)}{\partial r} = \mathcal{O}\left(\frac{1}{r^2}\right) \tag{13.1.20}$$

as $r \to \infty$, uniformly in θ. The function $f \in C(S)$ is assumed to satisfy

$$\int_S f(Q)\, dS = 0 \tag{13.1.21}$$

just as in (13.1.3) for the interior Neumann problem.

Combining (13.1.18) with (13.1.21) yields

$$\int_{\widehat{S}} \widehat{f}(\xi,\eta)\, d\widehat{S} = 0. \tag{13.1.22}$$

The problem (13.1.19) converts to the equivalent interior problem of finding $\widehat{u}$ satisfying

$$\begin{aligned} \Delta\widehat{u}(\xi,\eta) &= 0, && (\xi,\eta)\in\widehat{D},\\ \frac{\partial\widehat{u}(\xi,\eta)}{\partial\widehat{\boldsymbol{n}}(\xi,\eta)} &= \widehat{f}(\xi,\eta), && (\xi,\eta)\in\widehat{S},\\ \widehat{u}(0,0) &= 0. \end{aligned} \tag{13.1.23}$$

By Theorem 13.1.31 and (13.1.22), this has a unique solution $\widehat{u}$. This gives a complete solvability theory for the exterior Neumann problem.

The converted problems (13.1.15) and (13.1.23) can also be used for numerical purposes, and later we will return to these reformulations of exterior problems for Laplace's equation.

Green's representation formula on exterior regions

From the form of solutions to the interior Dirichlet problem, and using the Kelvin transform, we can assume the following form for potential functions u defined on D_e:

$$u(r\cos\theta, r\sin\theta) = u(\infty) + \frac{c(\theta)}{r} + \mathcal{O}\left(\frac{1}{r^2}\right) \tag{13.1.24}$$

as $r\to\infty$ and with $c(\theta) = A\cos\theta + B\sin\theta$ for suitable constants A, B. The notation $u(\infty)$ denotes the limiting value of $u(r\cos\theta, r\sin\theta)$ as $r\to\infty$. From this, we can use the Green's representation formulas (13.1.7)–(13.1.9) for interior regions to obtain the following Green's representation formula for potential functions on exterior regions.

$$\begin{aligned} u(A) = u(\infty) &- \frac{1}{2\pi}\int_S \frac{\partial u(Q)}{\partial \boldsymbol{n}_Q}\log|A-Q|\,dS_Q\\ &+ \frac{1}{2\pi}\int_S u(Q)\frac{\partial}{\partial \boldsymbol{n}_Q}\log|A-Q|\,dS_Q, \quad A\in D_e. \end{aligned} \tag{13.1.25}$$

To obtain a limiting value as $A\to P\in S$, we need the limit

$$\begin{aligned} &\lim_{\substack{A\to P\\ A\in D_e}}\int_S u(Q)\frac{\partial}{\partial \boldsymbol{n}_Q}\log|A-Q|\,dS_Q\\ &\qquad = \pi u(P) + \int_S u(Q)\frac{\partial}{\partial \boldsymbol{n}_Q}\log|P-Q|\,dS_Q. \end{aligned} \tag{13.1.26}$$

Note the change of sign of $u(P)$ when compared to (13.1.8). Using this in (13.1.25), we obtain

$$\begin{aligned} u(P) = 2u(\infty) &- \frac{1}{\pi}\int_S \frac{\partial u(Q)}{\partial \boldsymbol{n}_Q}\log|P-Q|\,dS_Q\\ &+ \frac{1}{\pi}\int_S u(Q)\frac{\partial}{\partial \boldsymbol{n}_Q}\log|P-Q|\,dS_Q, \quad P\in S. \end{aligned} \tag{13.1.27}$$

13.1.3 Boundary integral equations of direct type

The equations (13.1.7) and (13.1.25) give representations for functions harmonic in D_i and D_e, respectively, in terms of u and $\partial u/\partial \boldsymbol{n}$ on the boundary S of these regions. When given one of these boundary functions, the equations (13.1.9) and (13.1.27) can often be used to obtain the remaining boundary function. Numerical methods based on (13.1.9) and (13.1.27) are said to be of "direct type", as they find u or $\partial u/\partial \boldsymbol{n}$ on the boundary and these are quantities that are often of immediate physical interest. We will illustrate some of the possible BIE of direct type, leaving others as problems for the reader.

The interior Dirichlet problem (13.1.1)

The boundary condition is $u(P) = f(P)$ on S; and using it, (13.1.9) can be written as

$$\frac{1}{\pi}\int_S \rho(Q)\log|P-Q|\,dS_Q = g(P), \quad P \in S. \tag{13.1.28}$$

To emphasize the form of the equation, we have introduced

$$\rho(Q) \equiv \frac{\partial u(Q)}{\partial \boldsymbol{n}_Q}, \quad g(P) \equiv f(P) + \frac{1}{\pi}\int_S f(Q)\frac{\partial}{\partial \boldsymbol{n}_Q}\log|P-Q|\,dS_Q.$$

The equation (13.1.28) is of the first kind, and it is often used as the prototype for studying boundary integral equations of the first kind. In Section 13.3, we discuss the solution of (13.1.28) in greater detail.

The interior Neumann problem (13.1.2)

The boundary condition is $\partial u/\partial \boldsymbol{n} = f$ on S; and using it, we write (13.1.9) as

$$\begin{aligned} u(P) + \frac{1}{\pi}\int_S u(Q)\frac{\partial}{\partial \boldsymbol{n}_Q}\log|P-Q|\,dS_Q \\ = \frac{1}{\pi}\int_S f(Q)\log|P-Q|\,dS_Q, \qquad P \in S. \end{aligned} \tag{13.1.29}$$

This is an integral equation of the second kind. Unfortunately, it is not uniquely solvable; and this should not be surprising when given the lack of unique solvability for the Neumann problem itself. The homogeneous equation has $u \equiv 1$ as a solution, as can be seen by substituting the harmonic function $u \equiv 1$ into (13.1.9). The equation (13.1.29) is solvable if and only if the boundary function f satisfies the condition (13.1.3). The simplest way to deal with the lack of uniqueness in solving (13.1.29) is to introduce an additional condition such as

$$u(P^*) = 0$$

for some fixed point $P^* \in S$. This will lead to a unique solution for (13.1.29). Combine this with the discretization of the integral equation, to obtain a suitable numerical approximation for u. There are other ways of converting (13.1.29) to a uniquely solvable equation, and some of these are explored in [10]. However, there are preferable alternative ways to solve the interior Neumann problem. One of the simplest is simply to convert it to an equivalent exterior Neumann problem, using the Kelvin transform given earlier; and then use techniques for the exterior problem, such as the BIE given in (13.1.30) below.

The exterior Neumann problem (13.1.19)

The boundary condition is $\partial u/\partial \boldsymbol{n} = f$ on S, and u also satisfies $u(\infty) = 0$. Using this, (13.1.27) becomes

$$\begin{aligned} u(P) - \frac{1}{\pi}\int_S u(Q)\frac{\partial}{\partial \boldsymbol{n}_Q}\log|P-Q|\,dS_Q \\ = -\frac{1}{\pi}\int_S f(Q)\,\log|P-Q|\,dS_Q, \qquad P \in S. \end{aligned} \tag{13.1.30}$$

This equation is uniquely solvable, as will be discussed in greater detail below, following (13.2.3) in Section 13.2. This is considered a practical approach to solving the exterior Neumann problem, especially when one wants to find only the boundary data $u(P)$, $P \in S$. The numerical solution of the exterior Neumann problem using this approach is given following (13.2.24) in Section 13.2.

As above, we assume the boundary S is a smooth simple closed curve with a twice-continuously differentiable parametrization. More precisely, let S be parameterized by

$$\boldsymbol{r}(t) = (\xi(t), \eta(t)), \quad 0 \le t \le L \tag{13.1.31}$$

with $\boldsymbol{r} \in C^2[0, L]$ and $|\boldsymbol{r}'(t)| \neq 0$ for $0 \le t \le L$. We assume the parametrization traverses S in a counter-clockwise direction. We usually consider $\boldsymbol{r}(t)$ as being extended periodically from $[0, L]$ to $(-\infty, \infty)$; and we write $\boldsymbol{r} \in C_p^2(L)$, generalizing from the definition of $C_p^2(2\pi)$ given in Chapter 1. Introduce the interior unit normal $\boldsymbol{n}(t)$ which is orthogonal to the curve S at $\boldsymbol{r}(t)$:

$$\boldsymbol{n}(t) = \frac{(-\eta'(t), \xi'(t))}{\sqrt{\xi'(t)^2 + \eta'(t)^2}}.$$

Using this representation $\boldsymbol{r}(t)$ for S, and multiplying in (13.1.30) by $(-\pi)$, we can rewrite (13.1.30) as

$$-\pi u(t) + \int_0^L k(t,s)u(s)\,ds = g(t), \quad 0 \le t \le L \tag{13.1.32}$$

where

$$k(t,s) = \frac{\eta'(s)[\xi(t)-\xi(s)] - \xi'(s)[\eta(t)-\eta(s)]}{[\xi(t)-\xi(s)]^2 + [\eta(t)-\eta(s)]^2}$$
$$= \frac{\eta'(s)\xi[s,s,t] - \xi'(s)\eta[s,s,t]}{|\boldsymbol{r}[s,t]|^2}, \quad s \neq t, \tag{13.1.33}$$
$$k(t,t) = \frac{\eta'(t)\xi''(t) - \xi'(t)\eta''(t)}{2\,[\xi'(t)^2 + \eta'(t)^2]} \tag{13.1.34}$$

and

$$g(t) = \int_0^L f(\boldsymbol{r}(s))\sqrt{\xi'(s)^2 + \eta'(s)^2}\,\log|\boldsymbol{r}(t) - \boldsymbol{r}(s)|\;ds. \tag{13.1.35}$$

In (13.1.32), we have used $u(t) \equiv u(\boldsymbol{r}(t))$, for simplicity in notation. The second fraction in (13.1.33) uses first and second order Newton divided differences, to more easily obtain the limiting value $k(t,t)$ of (13.1.34). The value of $k(t,t)$ is one-half the curvature of S at $\boldsymbol{r}(t)$.

As in earlier chapters, we write (13.1.32) symbolically as

$$(-\pi + K)\,u = g. \tag{13.1.36}$$

By examining the formulas for $k(t,s)$, we have

$$\boldsymbol{r} \in C^{\kappa}[0,L] \implies k \in C^{\kappa-2}([0,L]\times[0,L]). \tag{13.1.37}$$

The kernel function k is periodic in both variables, with period L, as are also the functions u and g.

Recall from Example 1.2.28(a) of Chapter 1 the space $C_p^{\ell}(2\pi)$ of all ℓ-times continuously differentiable and periodic functions on $(-\infty,\infty)$. Since the parameterization $\boldsymbol{r}(t)$ is on $[0,L]$, we generalize $C_p^{\ell}(2\pi)$ to $C_p^{\ell}(L)$, with functions having period L on $(-\infty,\infty)$. The norm is

$$\|h\|_{\ell} = \max\left\{\|h\|_{\infty}, \|h'\|_{\infty}, \ldots, \|h^{(\ell)}\|_{\infty}\right\}$$

with the maximum norm taken over the interval $[0,L]$. We always assume for the parametrization that $\boldsymbol{r} \in C_p^{\kappa}(L)$, with $\kappa \geq 2$; and therefore the integral operator K is a compact operator from $C_p(L)$ to $C_p(L)$. Moreover, from (13.1.37), K maps $C_p(L)$ to $C_p^{\kappa-2}(L)$. The numerical solution of (13.1.32) is examined in detail in Section 13.2, along with related integral equations.

Using the Kelvin transform, the interior Neumann problem (13.1.1) can be converted to an equivalent exterior Neumann problem, as was done in passing between (13.1.19) and (13.1.23). Solving the exterior problem will correspond to finding that solution to the interior Neumann problem which is zero at the origin (where we assume $\mathbf{0} \in D_i$).

The exterior Dirichlet problem (13.1.14)

The Kelvin transform can also be used to convert the exterior Dirichlet problem (13.1.14) to an equivalent interior Dirichlet problem. After doing so, there are many options for solving the interior problem, including using the first kind boundary integral equation (13.1.28). The value of $u(\infty)$ can be obtained as the value at $\mathbf{0}$ of the transformed problem.

Boundary integral equations of indirect type

Indirect BIE methods are based on representing the unknown harmonic function u as either a *single layer potential*,

$$u(A) = \int_S \rho(Q) \log|A - Q| \, dS_Q, \quad A \in \mathbb{R}^2, \tag{13.1.38}$$

or a *double layer potential,*

$$u(A) = \int_S \rho(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|A - Q| \, dS_Q, \quad A \in \mathbb{R}^2. \tag{13.1.39}$$

These have physical interpretations, for example, letting ρ denote a given *charge density* on S or a *dipole charge density* on S. For a classical interpretation of such potentials, see Kellogg [141]. Both of these formulas satisfy Laplace's equation for $A \in \mathbb{R}^2 \backslash S$. The density ρ is to be chosen such that u satisfies given boundary conditions on S.

Double layer potentials

Suppose the function u is the solution of the interior Dirichlet problem with $u \equiv f$ on S. Then use (13.1.8) to take limits in (13.1.39) as $A \to P \in S$. This yields the boundary integral equation

$$-\pi\rho(P) + \int_S \rho(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|P - Q| \, dS_Q = f(P), \quad P \in S. \tag{13.1.40}$$

Note that the form of the left side of this equation is exactly that of (13.1.32) for the exterior Neumann problem. We discuss in detail the numerical solution of this and related equations in Section 13.2. Ivar Fredholm used (13.1.39) to show the solvability of the interior Dirichlet problem for Laplace's equation, and he did so by showing (13.1.40) is uniquely solvable for all $f \in C(S)$.

The use of (13.1.40) gives a BIE of "indirect type", as the solution ρ is usually of only indirect interest, it being a means of obtaining u using (13.1.39). Usually, ρ has no immediate physical significance.

Single layer potentials

The single layer potentials are also used to solve interior and exterior problems, for both Dirichlet and Neumann problems. The single layer potential (13.1.38) satisfies Laplace's equation in $D_i \cup D_e$, and it is continuous

in $\mathbb{R}^2$, provided $\rho \in L^1(S)$. For example, to solve the interior Dirichlet problem with boundary data $u = f$ on S, we must solve the first kind integral equation

$$\int_S \rho(Q)\, \log|P - Q|\, dS_Q = f(P), \quad P \in S. \tag{13.1.41}$$

Some additional properties of the single layer potential are examined in the exercises at the end of this section.

If we seek the solution of the interior Neumann problem (13.1.2) as a single layer potential (13.1.38), with boundary data f on S, then the density ρ must satisfy

$$\pi\rho(P) + \int_S \rho(Q) \frac{\partial}{\partial \boldsymbol{n}_P} \log|P - Q|\, dS_Q = f(P), \quad P \in S. \tag{13.1.42}$$

To obtain this, begin by forming the normal derivative of (13.1.38),

$$\frac{\partial u(A)}{\partial \boldsymbol{n}_P} = \boldsymbol{n}_P \cdot \nabla_A \left[\int_S \rho(Q)\, \log|A - Q|\, dS_Q \right], \quad A \in D_i,\ \in S.$$

Take the limit as $A \to P \in S$. Using an argument similar to that used in obtaining (13.1.8), and applying the boundary condition $\partial u/\partial \boldsymbol{n}_P = f$, we obtain (13.1.42). The integral operator in (13.1.42) is the adjoint to that in (13.1.40); and the left side of the integral equation is the adjoint of the left side of (13.1.29).

The adjoint equation (13.1.29) is not uniquely solvable, as $\rho \equiv 1$ is a solution of the homogeneous equation. To see this, let $u \equiv 1$ (and $f \equiv 0$) in (13.1.29), thus showing that (13.1.29) is not a uniquely solvable equation. Since this is the adjoint equation to the homogeneous form of (13.1.42), we have that the latter is also not uniquely solvable (Theorem 2.8.14 in Subsection 2.8.5). An examination of how to obtain uniquely solvable variants of (13.1.42) is given in [10].

The single and double layer potentials of (13.1.38)–(13.1.39) can be given additional meaning by using the Green's representation formulas of this section. The density ρ can be related to the difference on the boundary S of solutions or their normal derivatives for Laplace's equation on the regions that are interior and exterior to S; see [18, pp. 317–320].

There are also additional representation formulas and boundary integral equations which can be obtained by other means. For example, representation formulas can be obtained from the Cauchy integral formula for functions of a complex variable. All analytic functions $f(z)$ can be written in the form

$$f(z) = u(x, y) + i\, v(x, y).$$

Using the Cauchy-Riemann equations for u and v, it follows that both u and v are harmonic functions in the domain of analyticity for f. For results

obtained from this approach, see Mikhlin [171]. Most of the representation formulas and BIE given in this section can also be obtained by using Cauchy's integral formula.

Exercise 13.1.1 Derive (13.1.5)–(13.1.6).

Exercise 13.1.2 Using (13.1.5), show that if the interior Dirichlet problem (13.1.1) has a solution, then it is unique.

Exercise 13.1.3 Derive (13.1.7), using the ideas sketched preceding the formula.

Exercise 13.1.4 Derive (13.1.11) and (13.1.13).

Exercise 13.1.5 Assume S is a smooth simple closed curve ($\boldsymbol{r} \in C_p^2(L)$). Prove

$$\int_S \log|A-Q|\, dS_Q = 2\pi, \quad A \in D.$$

What is the value of this integral if $A \in S$? If $A \in D_e$?

Exercise 13.1.6 Assume S is a smooth simple closed curve ($\boldsymbol{r} \in C_p^2(L)$). What are the values of

$$\int_S \frac{\partial}{\partial \boldsymbol{n}_Q} \log|A-Q|\, dS_Q$$

for the three cases of $A \in D$, $A \in S$, and $A \in D_e$?

Exercise 13.1.7 Derive the formulas given in (13.1.33)–(13.1.34), and then show (13.1.37).

Exercise 13.1.8 Consider the single layer potential u of (13.1.38). Show that

$$u(A) \approx c \log|A| \quad \text{as } |A| \to \infty.$$

What is c? Suppose you are solving the exterior Dirichlet problem by representing it as the single layer potential in (13.1.38), say with boundary data f on S. Then the density function ρ must satisfy the integral equation

$$\int_S \rho(Q) \log|P-Q|\, dS_Q = f(P), \quad P \in S.$$

In order to assure that this single layer potential u represents a function bounded at ∞, what additional condition must be imposed on the density function ρ?

Exercise 13.1.9 Derive the analogue of (13.1.33)–(13.1.34) for the integral operator in (13.1.42).

Exercise 13.1.10 Let the boundary parameterization for S be

$$\boldsymbol{r}(t) = \gamma(t)\,(\cos t, \sin t), \quad 0 \le t \le 2\pi$$

with $\gamma(t)$ a twice continuously and positive 2π-periodic function on $[0, 2\pi]$. Find the kernel function $k(t,s)$ of (13.1.32)–(13.1.34) for this boundary, simplifying as much as possible. What happens when $s - t \to 0$?

Exercise 13.1.11 Generalize the preceding Exercise 13.1.10 to the boundary parameterization

$$\boldsymbol{r}(t) = \gamma(t)\,(a\cos t, b\sin t), \quad 0 \le t \le 2\pi$$

with $a, b > 0$, and $\gamma(t)$ a twice continuously and positive 2π-periodic function on $[0, 2\pi]$. Find the kernel function $k(t, s)$ of (13.1.32)–(13.1.34) for this boundary, simplifying as much as possible.

13.2 Boundary integral equations of the second kind

The original theory developed by Ivar Fredholm for the solvability of integral equations was for the boundary integral equations of the second kind introduced in the preceding section; and these equations have also long been used as a means to solve boundary value problems for Laplace's equation. In this section, we consider the numerical solution of these boundary integral equations of the second kind. We begin with a classic indirect method for solving the interior Dirichlet problem for Laplace's equation; and then the results for this method are extended to integral equations for the interior and exterior Neumann problems.

Recall the double layer representation (13.1.39) for a function u harmonic on the interior region D_i:

$$u(A) = \int_S \rho(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|A - Q|\, dS_Q, \quad A \in D_i. \tag{13.2.1}$$

To solve the interior Dirichlet problem (13.1.1), the density ρ is obtained by solving the boundary integral equation given in (13.1.40), namely

$$-\pi\rho(P) + \int_S \rho(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|P - Q|\, dS_Q = f(P), \quad P \in S \tag{13.2.2}$$

with f the given value of u on S. This is basically the same form of integral equation as in (13.1.30) for the exterior Neumann problem, with a different right hand function. When the representation $\boldsymbol{r}(t) = (\xi(t), \eta(t))$ of (13.1.31) for S is applied, this integral equation becomes

$$-\pi\rho(t) + \int_0^L k(t, s)\rho(s)\, ds = f(t), \quad 0 \le t \le L \tag{13.2.3}$$

with $k(t, s)$ given in (13.1.33)–(13.1.34) and $f(t) \equiv f(\boldsymbol{r}(t))$. The smoothness and periodicity of k is discussed in and following (13.1.37); and the natural function space setting for studying (13.2.3) is $C_p(L)$ with the uniform norm. Symbolically, we write (13.2.3) as $(-\pi + K)\rho = f$.

The equation (13.2.2) has been very well studied, for over a century; for example, see the references and discussion of this equation in Colton [57, p. 216], Kress [149, p. 71] and Mikhlin [171, Chap. 4]. From this work, $(-\pi + K)^{-1}$ exists as a bounded operator from $C_p(L)$ to $C_p(L)$.

The functions $f, \rho \in C_p(L)$, and the kernel k is periodic in both variables, with period L, over $(-\infty, \infty)$; and in addition, both k and ρ are usually smooth functions. Thus the most efficient numerical method for solving the equation (13.2.3) is generally the Nyström method with the trapezoidal rule as the numerical integration rule. Recall the extensive discussion of the trapezoidal rule in Proposition 7.5.6 of Chapter 7.

Because of the periodicity, the trapezoidal rule simplifies further, and the approximating equation takes the form

$$-\pi\rho_n(t) + h\sum_{j=1}^{n} k(t,t_j)\rho_n(t_j) = f(t), \quad 0 \le t \le L \tag{13.2.4}$$

with $h = L/n$, $t_j = jh$ for $j = 1, 2, \ldots, n$. Symbolically, we write this as $(-\pi + K_n)\rho_n = f$, with the numerical integration operator K_n defined implicitly by (13.2.4). Collocating at the node points, we obtain the linear system

$$-\pi\rho_n(t_i) + h\sum_{j=1}^{n} k(t_i,t_j)\rho_n(t_j) = f(t_i), \quad i = 1, \ldots, n \tag{13.2.5}$$

whose solution is $(\rho_n(t_1), \ldots, \rho_n(t_n))^T$. Then the Nyström interpolation formula can be used to obtain $\rho_n(t)$:

$$\rho_n(t) = \frac{1}{\pi}\left[-f(t) + h\sum_{j=1}^{n} k(t,t_j)\rho_n(t_j)\right], \quad 0 \le t \le L. \tag{13.2.6}$$

This is a simple method to program; and usually the value of n is not too large, so that the linear system (13.2.5) can be solved directly, without iteration.

The error analysis for the above is straightforward from Theorem 12.4.4 of Chapter 12. This theorem shows that (13.2.4) is uniquely solvable for all sufficiently large values of n, say $n \ge N$; and moreover,

$$\|\rho - \rho_n\|_\infty \le \left\|(-\pi + K_n)^{-1}\right\| \|K\rho - K_n\rho\|_\infty, \quad n \ge N. \tag{13.2.7}$$

It is well known that the trapezoidal rule is very rapidly convergent when the integrand is periodic and smooth; and consequently, $\rho_n \to \rho$ with a similarly rapid rate of convergence.

Example 13.2.1 Let the boundary S be the ellipse

$$\boldsymbol{r}(t) = (a\cos t, b\sin t), \quad 0 \le t \le 2\pi. \tag{13.2.8}$$

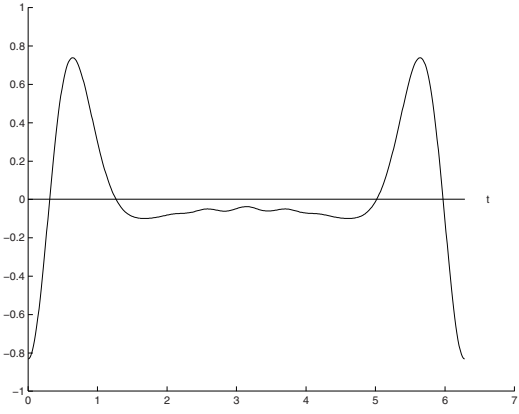

FIGURE 13.1. The density ρ for (13.2.10) with $(a, b) = (1, 5)$

n	$(a,b)=(1,2)$	$(a,b)=(1,5)$	$(a,b)=(1,8)$
8	3.67E−3	4.42E−1	3.67E+0
16	5.75E−5	1.13E−2	1.47E−1
32	1.34E−14	1.74E−5	1.84E−3
64		3.96E−11	6.66E−7
128			7.23E−14

TABLE 13.1. Errors in density function ρ_n for (13.2.10)

In this case, the kernel k of (13.1.33) can be reduced to

$$k(t,s) = \kappa\left(\frac{s+t}{2}\right), \quad \kappa(\theta) = \frac{-ab}{2\left(a^2\sin^2\theta + b^2\cos^2\theta\right)} \tag{13.2.9}$$

and the integral equation (13.2.3) becomes

$$-\pi\rho(t) + \int_0^{2\pi} \kappa\left(\frac{s+t}{2}\right)\rho(s)\,ds = f(t), \quad 0 \le t \le 2\pi. \tag{13.2.10}$$

In Table 13.1, we give results for solving this equation with

$$f(x,y) = e^x \cos y, \quad (x,y) \in S. \tag{13.2.11}$$

The true solution ρ is not known explicitly; but we obtain a highly accurate solution by using a large value of n, and then this solution is used to calculate the errors shown in the table.

Results are given for $(a, b) = (1, 2)$ and $(1, 5)$. The latter ellipse is somewhat elongated, and this causes the kernel k to be more peaked. In partic-

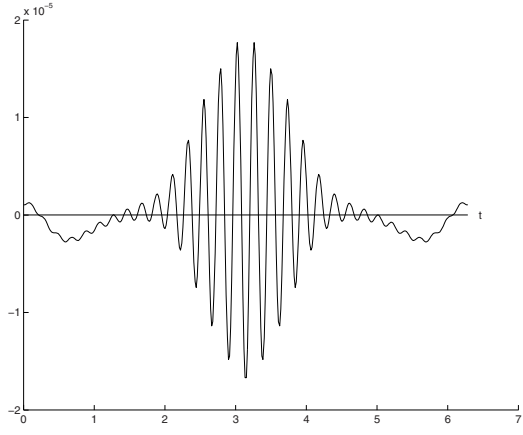

FIGURE 13.2. The error $\rho - \rho_{32}$ for (13.2.10) with $(a, b) = (1, 5)$.

ular, introduce the *peaking factor*

$$p(a,b) \equiv \frac{\max|k(t,s)|}{\min|k(t,s)|} = \left(\frac{\max\{a,b\}}{\min\{a,b\}}\right)^2.$$

Then $p(1,2) = 4$, $p(1,5) = 25$, $p(1,8) = 64$. As the peaking factor becomes larger, it is necessary to increase n in order to retain comparable accuracy in approximating the integral $K\rho$, and the consequences of this can be seen in the table.

A graph of ρ is given in Figure 13.1 for $(a, b) = (1, 5)$, and it shows a somewhat rapid change in the function around $t = 0$ or $(x, y) = (a, 0)$ on S. For the same curve S, a graph of the error $\rho(t) - \rho_n(t)$, $0 \le t \le 2\pi$, is given in Figure 13.2 for the case $n = 32$. Perhaps surprisingly in light of Figure 13.1, the error is largest around $t = \pi$ or $(x, y) = (-a, 0)$ on S, where ρ is better behaved. □

13.2.1 Evaluation of the double layer potential

When using the representation $\boldsymbol{r}(s) = (\xi(s), \eta(s))$ of (13.1.31) for S, the double layer integral formula (13.2.1) takes the form

$$u(x,y) = \int_0^L M(x,y,s)\rho(s)\,ds, \quad (x,y) \in D_i \tag{13.2.12}$$

where

$$M(x,y,s) = \frac{-\eta'(s)[\xi(s)-x] + \xi'(s)[\eta(s)-y]}{[\xi(s)-x]^2 + [\eta(s)-y]^2}. \tag{13.2.13}$$

This kernel is increasingly peaked as (x, y) approaches S. To see this more clearly, let S be the unit circle given by $\boldsymbol{r}(s) = (\cos s, \sin s)$, $0 \le s \le 2\pi$. Then

$$M(x, y, s) = \frac{-\cos s\,(\cos s - x) - \sin s\,(\sin s - y)}{(\cos s - x)^2 + (\sin s - y)^2}.$$

To see the near-singular behaviour more clearly, let (x, y) approach the point $(\cos s, \sin s)$ along the line

$$(x, y) = q(\cos s, \sin s), \quad 0 \le q < 1.$$

Then after simplifying,

$$M(q\cos s, q\sin s, s) = \frac{1}{1-q}.$$

The integrand of (13.2.12) is increasingly peaked as $q \to 1-$.

We use numerical integration to approximate (13.2.12); and since the integrand is periodic in s, the trapezoidal rule is an optimal choice when choosing among regular quadrature rules with uniformly distributed quadrature nodes. As (x, y) approaches S, the needed number of integration nodes will need to be increased in order to retain equivalent accuracy in the approximate values of $u(x, y)$. For (x, y) very close to S, other means should be used to approximate the integral (13.2.12), since the trapezoidal rule will be very expensive.

To solve the original Dirichlet problem (13.1.1), we first approximate the density ρ, obtaining ρ_n; and then we numerically integrate the double layer integral based on ρ_n. To aid in studying the resulting approximation of $u(x, y)$, introduce the following notation. Let $u_n(x, y)$ be the double layer potential using the approximate density ρ_n obtained by the Nyström method of (13.2.4):

$$u_n(x, y) = \int_0^L M(x, y, s)\rho_n(s)\,ds, \quad (x, y) \in D_i. \tag{13.2.14}$$

Let $u_{n,m}(x, y)$ denote the result of approximating $u_n(x, y)$ using the trapezoidal rule:

$$u_{n,m}(x, y) = h\sum_{i=1}^{m} M(x, y, t_i)\rho_n(t_i), \quad (x, y) \in D_i. \tag{13.2.15}$$

For the error in u_n, note that $u - u_n$ is a harmonic function; and therefore, by the maximum principle for such functions,

$$\max_{(x,y)\in\overline{D}_i} |u(x, y) - u_n(x, y)| = \max_{(x,y)\in S} |u(x, y) - u_n(x, y)|. \tag{13.2.16}$$

Since $u - u_n$ is also a double layer potential, the argument which led to the original integral equation (13.2.1) also implies

$$u(P) - u_n(P) = -\pi[\rho(P) - \rho_n(P)] + \int_S [\rho(Q) - \rho_n(Q)] \frac{\partial}{\partial \boldsymbol{n}_Q} \log|P - Q| \, dS_Q, \quad P \in S. \tag{13.2.17}$$

Taking bounds,

$$|u(P) - u_n(P)| \leq (\pi + \|K\|) \|\rho - \rho_n\|_\infty, \quad P \in S. \tag{13.2.18}$$

Combined with (13.2.16),

$$\max_{(x,y)\in\overline{D}_i} |u(x,y) - u_n(x,y)| \leq (\pi + \|K\|) \|\rho - \rho_n\|_\infty. \tag{13.2.19}$$

If the region D_i is convex, then the double layer kernel is strictly negative; and it can then be shown that

$$\|K\| = \pi. \tag{13.2.20}$$

For convex regions, therefore,

$$\max_{(x,y)\in\overline{D}_i} |u(x,y) - u_n(x,y)| \leq 2\pi \|\rho - \rho_n\|_\infty. \tag{13.2.21}$$

An algorithm for solving the interior Dirichlet problem (13.1.1) can be based on first solving for ρ_n to a prescribed accuracy. Then (13.2.19) says u_n has comparable accuracy uniformly on D_i. To complete the task of evaluating $u_n(x,y)$ for given values of (x,y), one can use the trapezoidal rule (13.2.15), varying m to obtain desired accuracy in $u_{n,m}(x,y)$. The total error is then given by

$$u(x,y) - u_{n,m}(x,y) = [u(x,y) - u_n(x,y)] + [u_n(x,y) - u_{n,m}(x,y)]. \tag{13.2.22}$$

Ideally, the two errors on the right side should be made comparable in size, to make the algorithm as efficient as possible. A Fortran program implementing these ideas is given in [24], and it also uses a slight improvement on (13.2.15) when (x,y) is near to S.

Example 13.2.2 We continue with Example 13.2.1, noting that the true solution is also given by (13.2.11). For the case $(a,b) = (1,5)$ and $n = 32$, we examine the error in the numerical solutions u_n and $u_{n,m}$ along the line

$$\boldsymbol{c}(q) = q(a\cos\frac{\pi}{4}, b\sin\frac{\pi}{4}), \quad 0 \leq q < 1. \tag{13.2.23}$$

A graph of the error $u(\boldsymbol{c}(q)) - u_n(\boldsymbol{c}(q))$, $0 \leq q \leq .94$, is shown in Figure 13.3. Note that the size of the error is around 20 times smaller than is

r	$m = 32$	$m = 64$	$m = 128$	$m = 256$
0	$-1.34E-2$	$-2.20E-5$	$-1.68E-6$	$-1.68E-6$
.20	$1.60E-2$	$6.89E-5$	$-1.82E-6$	$-1.82E-6$
.40	$1.14E-3$	$1.94E-5$	$-1.49E-7$	$-1.58E-7$
.60	$-7.88E-2$	$-3.5E-3$	$4.63E-6$	$2.22E-6$
.80	$5.28E-1$	$2.33E-2$	$-1.31E-3$	$4.71E-6$
.90	$-1.13E+0$	$4.82E-1$	$3.12E-2$	$-2.64E-4$
.94	$-1.08E+0$	$-8.44E-1$	$2.05E-1$	$1.85E-3$

TABLE 13.2. Errors $u(\boldsymbol{c}(q)) - u_{n,m}(\boldsymbol{c}(q))$ with $n = 32$

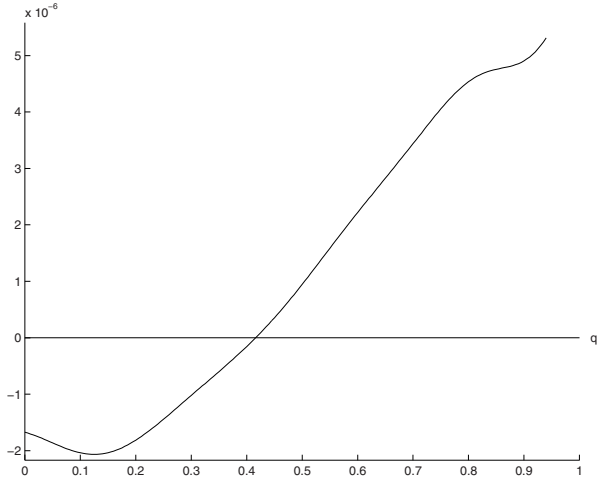

FIGURE 13.3. The errors $u(\boldsymbol{c}(q)) - u_n(\boldsymbol{c}(q))$ with $n = 32$

predicted from the error of $\|\rho - \rho_{32}\|_\infty = 1.74 \times 10^{-5}$ of Table 13.1 and the bound (13.2.21). Table 13.2 contains the errors $u(\boldsymbol{c}(q)) - u_{n,m}(\boldsymbol{c}(q))$ for selected values of q and m, with $n = 32$. Graphs of these errors are given in Figure 13.4. Compare these graphs with that of Figure 13.3, noting the quite different vertical scales. It is clear that increasing m decreases the error, up to the point that the dominant error is that of $u(x, y) - u_n(x, y)$ in (13.2.22). □

13.2.2 The exterior Neumann problem

Recall the solving of the exterior Neumann problem (13.1.2) by means of the integral representation formula (13.1.25) and the boundary integral

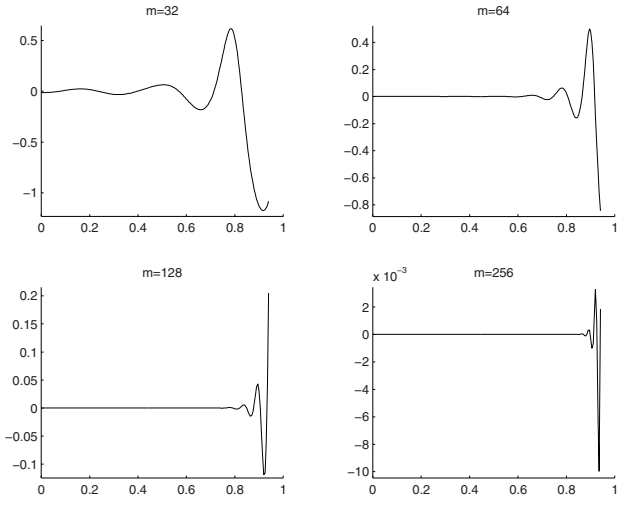

FIGURE 13.4. The errors $u(\boldsymbol{c}(q)) - u_{n,m}(\boldsymbol{c}(q))$ with $n = 32$.

equation of (13.1.30). We rewrite the latter as

$$\begin{aligned} -\pi u(P) + \int_S u(Q) \frac{\partial}{\partial \boldsymbol{n}_Q} \log|P-Q| \, dS_Q \\ = \int_S f(Q) \log|P-Q| \, dS_Q, \qquad P \in S. \end{aligned} \tag{13.2.24}$$

The left side of this equation is the same as that of (13.2.2) for the interior Dirichlet problem; and it is therefore only the evaluation of the right side which concerns us here. Recalling (13.1.35), the right side is

$$g(t) = \int_0^L f(\boldsymbol{r}(s)) \sqrt{\xi'(s)^2 + \eta'(s)^2} \log|\boldsymbol{r}(t) - \boldsymbol{r}(s)| \, ds. \tag{13.2.25}$$

This could be approximated using the product integration techniques of Section 12.5 in Chapter 12; but we consider a more efficient method.

To simplify the notation, the parametrization $\boldsymbol{r}(t)$ of (13.2.1) is assumed to be defined on the standard interval $[0, 2\pi]$. Also, introduce

$$\varphi(s) = f(\boldsymbol{r}(s)) \sqrt{\xi'(s)^2 + \eta'(s)^2}. \tag{13.2.26}$$

The integral (13.2.25) becomes

$$g(t) = \int_0^{2\pi} \varphi(s) \log|\boldsymbol{r}(t) - \boldsymbol{r}(s)| \, ds, \qquad 0 \le t \le 2\pi. \tag{13.2.27}$$

We write the kernel of this integral in the form

$$\log|\boldsymbol{r}(t)-\boldsymbol{r}(s)| = \log\left|2e^{-1/2}\sin\left(\frac{t-s}{2}\right)\right| - \pi b(t,s) \tag{13.2.28}$$

with

$$b(t,s) = \begin{cases} -\dfrac{1}{\pi}\log\dfrac{\left|e^{1/2}[\boldsymbol{r}(t)-\boldsymbol{r}(s)]\right|}{\left|2\sin\left(\dfrac{t-s}{2}\right)\right|}, & t-s\neq 2m\pi, \\ -\dfrac{1}{\pi}\log\left|e^{1/2}\boldsymbol{r}'(t)\right|, & t-s=2m\pi. \end{cases} \tag{13.2.29}$$

The integral (13.2.27) becomes

$$\begin{aligned} g(t) &= -\pi\left[-\frac{1}{\pi}\int_0^{2\pi}\varphi(s)\log\left|2e^{-1/2}\sin\left(\frac{t-s}{2}\right)\right|ds + \int_0^{2\pi} b(t,s)\varphi(s)\,ds\right] \\ &\equiv -\pi[\mathcal{A}\varphi(t)+B\varphi(t)]. \end{aligned} \tag{13.2.30}$$

Assuming $\boldsymbol{r}\in C_p^{\kappa}(2\pi)$, the kernel function $b\in C^{\kappa-1}([0,2\pi]\times[0,2\pi])$; and b is periodic in both variables t and s. Consequently, the second integral $B\varphi(t)$ in (13.2.30) can be accurately and efficiently approximated using the trapezoidal rule.

The first integral in (13.2.30) is a minor modification of the integral operator associated with the kernel $\log|P-Q|$ for S equal to the unit circle about the origin, where we have

$$\mathcal{A}\varphi(t) = -\frac{1}{\pi}\int_0^{2\pi}\varphi(s)\log\left|2e^{-1/2}\sin\left(\frac{t-s}{2}\right)\right|ds, \quad 0\le t\le 2\pi. \tag{13.2.31}$$

This operator was introduced in Section 7.5.4 and some properties of it were given there.

In particular,

$$\mathcal{A}\varphi(t) = \frac{1}{\sqrt{2\pi}}\left[\widehat{\varphi}(0) + \sum_{|m|>0}\frac{\widehat{\varphi}(m)}{|m|}e^{imt}\right], \tag{13.2.32}$$

based on the Fourier series

$$\varphi(s) = \frac{1}{\sqrt{2\pi}}\sum_{m=-\infty}^{\infty}\widehat{\varphi}(m)e^{ims}$$

for an arbitrary $\varphi\in L^2(0,2\pi)$. This is an expansion of $\mathcal{A}\varphi$ using the eigenfunctions $\psi_m(t)\equiv e^{imt}$ and the corresponding eigenvalues of $\mathcal{A}$. For a proof of this result, and for a much more extensive discussion of the properties of $\mathcal{A}$, see Yan and Sloan [242].

As noted in Section 7.5.4, (13.2.32) can be used to show that $\mathcal{A}$ is a bijective bounded linear operator from $H^0(2\pi) \equiv L^2(0, 2\pi)$ to $H^1(2\pi)$, with $\|\mathcal{A}\| = 1$ for this mapping. The Sobolev space $H^1(2\pi)$ was introduced in Definition 7.5.1 of Chapter 7. When $\mathcal{A}$ is considered as an operator from $C_p(2\pi)$ to $C_p(2\pi)$, we can show

$$\|\mathcal{A}\| \leq \sqrt{1 + \frac{\pi^2}{3}} \doteq 2.07. \tag{13.2.33}$$

For a derivation of this last bound, see [18, p. 330].

To approximate $\mathcal{A}\varphi$, we approximate φ using trigonometric interpolation; and then (13.2.32) is used to evaluate exactly the resulting approximation of $\mathcal{A}\varphi$. Let $n \geq 1$, $h = 2\pi/(2n+1)$, and

$$t_j = jh, \quad j = 0, \pm 1, \pm 2, \ldots \tag{13.2.34}$$

Let $\mathcal{Q}_n\varphi$ denote the trigonometric polynomial of degree $\leq n$ which interpolates $\varphi(t)$ at the nodes $\{t_0, t_1, \ldots, t_{2n}\}$, and by periodicity at all other nodes t_j (Theorem 7.5.7 of Chapter 7). Also, let $T_k(\varphi)$ denote the trapezoidal rule on $[0, 2\pi]$ with k subdivisions:

$$T_k(\varphi) = \frac{2\pi}{k} \sum_{j=0}^{k-1} \varphi\left(\frac{2\pi j}{k}\right), \quad \varphi \in C_p(2\pi).$$

The interpolation polynomial can be written as

$$\mathcal{Q}_n\varphi(t) = \sum_{j=-n}^{n} \alpha_j e^{ijt}. \tag{13.2.35}$$

The coefficients $\{\alpha_j\}$ can be obtained as numerical quadratures of the standard Fourier coefficients of φ; see [18, p. 331].

For the error in $\mathcal{Q}_n\varphi$, recall the error bound (3.7.21) in Chapter 3. Then

$$\|\varphi - \mathcal{Q}_n\varphi\|_\infty = \mathcal{O}\left(\frac{\log n}{n^{\ell+\alpha}}\right), \quad \varphi \in C_p^{\ell,\alpha}(2\pi). \tag{13.2.36}$$

In this relation, φ is assumed to be ℓ-times continuously differentiable, and $\varphi^{(\ell)}$ is assumed to satisfy the Hölder condition

$$\left|\varphi^{(\ell)}(s) - \varphi^{(\ell)}(t)\right| \leq c\,|s-t|^\alpha, \quad -\infty < s, t < \infty$$

with c a finite constant.

We approximate $\mathcal{A}\varphi(t)$ using $\mathcal{A}\mathcal{Q}_n\varphi(t)$. From (13.2.32),

$$\mathcal{A}\psi_j = \begin{cases} 1, & j = 0, \\ \dfrac{1}{|j|} e^{ijt}, & |j| > 0. \end{cases} \tag{13.2.37}$$

Applying this with (13.2.35),

$$\mathcal{A}\varphi(t) \approx \mathcal{A}\mathcal{Q}_n\varphi(t) = \alpha_0 + \sum_{\substack{j=-n \\ j\neq 0}}^{n} \frac{\alpha_j}{|j|} e^{ijt}, \quad -\infty < t < \infty. \tag{13.2.38}$$

To bound the error in $\mathcal{A}\mathcal{Q}_n\varphi$, we apply (13.2.33), yielding

$$\|\mathcal{A}\varphi - \mathcal{A}\mathcal{Q}_n\varphi\|_\infty \leq \|\mathcal{A}\| \, \|\varphi - \mathcal{Q}_n\varphi\|_\infty .$$

Using (13.2.36), this bound implies

$$\|\mathcal{A}\varphi - \mathcal{A}\mathcal{Q}_n\varphi\|_\infty = \mathcal{O}\left(\frac{\log n}{n^{\ell+\alpha}}\right), \quad \varphi \in C_p^{\ell,\alpha}(2\pi) \tag{13.2.39}$$

provided $\ell + \alpha > 0$. The approximation $\mathcal{A}\mathcal{Q}_n\varphi$ is rapidly convergent to $\mathcal{A}\varphi$.

To complete the approximation of the original integral (13.2.30), approximate $B\varphi(t)$ using the trapezoidal rule with the nodes $\{t_j\}$ of (13.2.34):

$$\begin{aligned} B\varphi(t) &\approx T_{2n+1}(b(t,\cdot)\varphi) \\ &= \frac{2\pi}{2n+1} \sum_{k=0}^{2n} b(t,t_k)\varphi(t_k) \\ &\equiv B_n\varphi(t). \end{aligned} \tag{13.2.40}$$

To bound the error, we can use the standard Euler-MacLaurin error formula [18, p. 285] to show

$$|B\varphi(t) - B_n\varphi(t)| \leq O(n^{-\ell}), \quad \varphi \in C_p^\ell(2\pi). \tag{13.2.41}$$

This assumes that $\boldsymbol{r} \in C_p^\kappa(2\pi)$ with $\kappa \geq \ell + 1$.

To solve the original integral equation (13.2.24), we use the Nyström method of (13.2.4)–(13.2.6) based on the trapezoidal numerical integration method with the $2n+1$ nodes $\{t_0, \ldots, t_{2n}\}$ of (13.2.34). The right side g of (13.2.30) is approximated by using (13.2.38) and (13.2.40), yielding the approximation

$$(-\pi + K_n)\, u_n = -\pi \left[\mathcal{A}\mathcal{Q}_n\varphi + B_n\varphi(t)\right]. \tag{13.2.42}$$

Error bounds can be produced by combining (13.2.39) and (13.2.41) with the earlier error analysis based on (13.2.7). We leave it as an exercise to show that if $\varphi \in C_p^\ell(2\pi)$ for some $\ell \geq 1$, and if $\boldsymbol{r} \in C_p^\kappa(2\pi)$ with $\kappa \geq \ell + 1$, then the approximate Nyström solution u_n of (13.2.24) satisfies

$$\|u - u_n\|_\infty \leq \mathcal{O}\left(\frac{\log n}{n^\ell}\right). \tag{13.2.43}$$

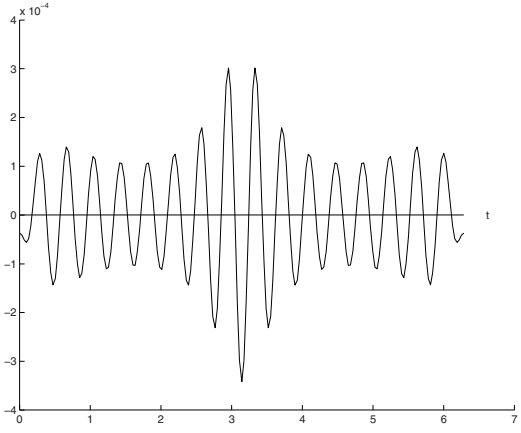

FIGURE 13.5. The error $u(\boldsymbol{r}(t)) - u_n(\boldsymbol{r}(t))$ for $n = 16$ and $(a, b) = (1, 2)$

Example 13.2.3 We solve the exterior Neumann problem on the region outside the ellipse

$$\boldsymbol{r}(t) = (a\cos t, b\sin t), \quad 0 \le t \le 2\pi.$$

For purposes of illustration, we use a known true solution,

$$u(x, y) = \frac{x}{x^2 + y^2}.$$

This function is harmonic; and $u(x, y) \to 0$ as $x^2 + y^2 \to \infty$. The Neumann boundary data is generated from u. Numerical results for $(a, b) = (1, 2)$ are given in Table 13.3; and in it, $m = 2n + 1$ is the order of the linear system being solved by the Nyström method. A graph of the error $u(\boldsymbol{r}(t)) - u_n(\boldsymbol{r}(t))$ is given in Figure 13.5 for the case $n = 16$. □

n	m	$\|u - u_n\|_\infty$
8	17	3.16E−2
16	33	3.42E−4
32	65	4.89E−8
64	129	1.44E−15

TABLE 13.3. The error $\|u - u_n\|_\infty$ for (13.2.42)

Exercise 13.2.1 Derive the integral equation (13.2.9)–(13.2.10) for solving the interior Dirichlet problem over an elliptical domain.

Exercise 13.2.2 Write a program to solve (13.2.10), implementing the Nyström method (13.2.4)–(13.2.6), as in Example 13.2.1. Experiment with varying values for n, (a, b), and boundary function f. For the latter, do experiments when f has a singular derivative (with respect to arc-length) on the boundary.

Exercise 13.2.3 Using the computation of Exercise 13.1.11, develop and program a numerical method for the parameterization

$$\boldsymbol{r}(t) = (2 + \cos t)\,(a\cos t, b\sin t), \quad 0 \le t \le 2\pi.$$

Do so for a variety of values of the positive constants a, b.

Exercise 13.2.4 Fill in the details of the arguments for the results given in (13.2.17)–(13.2.19).

Exercise 13.2.5 Prove that for D a bounded convex region, $\|K\| = \pi$, thus proving (13.2.20).

Exercise 13.2.6 Confirm the formulas (13.2.27)–(13.2.29).

Exercise 13.2.7 Assume $\boldsymbol{r} \in C_p^{\kappa}(2\pi)$. Show that the kernel function $b(t, s)$ of (13.2.29) belongs to $C^{\kappa-1}([0, 2\pi] \times [0, 2\pi])$ and that it is periodic in both variables t and s.

Exercise 13.2.8 Derive the results in the paragraph preceding (13.2.33). In particular, show $\mathcal{A}$ is a bijective mapping of $L^2(0, 2\pi)$ to $H^1(2\pi)$ with $\|\mathcal{A}\| = 1$ for this mapping.

13.3 A boundary integral equation of the first kind

Historically, most of the original theoretical work with boundary integral equations has been for integral equations of the second kind, and consequently, these types of boundary integral equations came to be the principal type used in applications. In addition, some integral equations of the first kind can be quite ill-conditioned, and this led some people to avoid such equations in general. Finally, numerical methods for integral equations of the first kind were difficult to analyze until somewhat recently.

Boundary integral equations of the first kind, however, are generally quite well-behaved; and recently, they have been an increasingly popular approach to solving various boundary value problems. In this section, we look at a well-studied boundary integral equation of the first kind, and we introduce some general analytical tools by means of doing an error analysis of a numerical method for solving this integral equation.

Returning to Section 13.1, the BIE (13.1.28) is an integral equation of the first kind of direct type. Introducing a change of sign, we write this integral equation as

$$-\frac{1}{\pi}\int_S \rho(Q)\log|P-Q|\,dS_Q = g(P), \quad P \in S. \tag{13.3.1}$$

In this case, the unknown density ρ is the value of the normal derivative on S of the unknown harmonic function u. This integral equation also arises as an indirect BIE for solving the interior Dirichlet problem for Laplace's equation (see (13.1.41) when $A \in S$). In this section, we consider various numerical methods for solving this integral equation, building on the ideas introduced in Section 13.2 following (13.2.25).

The solvability theory for (13.3.1) is well-developed, and an excellent presentation of it is given in Yan and Sloan [242]. In particular, if

$$\operatorname{diam}(D_i) < 1, \tag{13.3.2}$$

then the equation (13.3.1) is uniquely solvable for all $g \in H^1(S)$. The space $H^1(S)$ is equivalent to the space $H^1(2\pi)$ which was introduced in Definition 7.5.1 of Chapter 7, provided S is a smooth simple closed curve, as is assumed for this chapter. More generally, the integral equation is uniquely solvable if the equation

$$\int_S \psi(Q)\log|P-Q|\,dS_Q = 1, \quad P \in S \tag{13.3.3}$$

does not possess a solution. This is assured if (13.3.2) is satisfied; and since harmonic functions remain such under uniform scalar change of variables, we can assume (13.3.1) with no loss of generality. Curves S for which (13.3.3) has a solution are called "Γ-contours", and they are discussed at length in [242].

Write the first kind boundary integral equation (13.3.1) in the form

$$\begin{aligned} -\frac{1}{\pi}\int_0^{2\pi} \varphi(s)\log\left|2e^{-1/2}\sin\left(\frac{t-s}{2}\right)\right| ds \\ -\int_0^{2\pi} b(t,s)\varphi(s)\,ds = g(t), \quad 0 \le t \le 2\pi \end{aligned} \tag{13.3.4}$$

with $\varphi(s) \equiv \rho(\boldsymbol{r}(s))|\boldsymbol{r}'(s)|$. This decomposition of the integral operator of (13.3.1) was given earlier in (13.2.27)–(13.2.29). We write (13.3.4) in operator form as

$$\mathcal{A}\varphi + B\varphi = g. \tag{13.3.5}$$

Because of the continuity and smoothness properties of b, the operator B maps $H^q(2\pi)$ into $H^{q+2}(2\pi)$, at least. Using the embedding result that $H^{q+2}(2\pi)$ is compactly embedded in $H^{q+1}(2\pi)$ (Theorem 7.3.11), it follows

that B is a compact operator when considered as an operator from $H^q(2\pi)$ into $H^{q+1}(2\pi)$. Also, recall from (7.5.18) that

$$\mathcal{A} : H^q(2\pi) \underset{onto}{\overset{1-1}{\to}} H^{q+1}(2\pi), \quad q \geq 0. \tag{13.3.6}$$

On account of these mapping properties of $\mathcal{A}$ and B, we consider the integral equation (13.3.5) with the assumption $g \in H^{q+1}(2\pi)$; and we seek a solution $\varphi \in H^q(2\pi)$ to the equation.

From (13.3.6), the equation (13.3.5) is equivalent to

$$\varphi + \mathcal{A}^{-1}B\varphi = \mathcal{A}^{-1}g. \tag{13.3.7}$$

This is an integral equation of the second kind on $H^q(2\pi)$; and $\mathcal{A}^{-1}B$ is a compact integral operator when regarded as an operator on $H^q(2\pi)$ into itself. Consequently, the standard Fredholm alternative theorem applies; and if the homogeneous equation $\varphi+\mathcal{A}^{-1}B\varphi = 0$ has only the zero solution, then the original nonhomogeneous equation has a unique solution for all right sides $\mathcal{A}^{-1}g$. From [242], if S is not a Γ-contour, then the homogeneous version of the original integral equation (13.3.5) has only the zero solution; and thus by means of the Fredholm alternative theorem applied to (13.3.7), the integral equation (13.3.5) is uniquely solvable for all $g \in H^{q+1}(2\pi)$.

13.3.1 A numerical method

We give a numerical method for solving the first kind single layer equation (13.3.1) in the space $L^2(0, 2\pi)$. The method is a Galerkin method using trigonometric polynomials as approximations. We assume that the integral equation (13.3.1) is uniquely solvable for all $g \in H^1(2\pi)$.

For a given $n \geq 0$, introduce

$$V_n = \text{span}\{\psi_{-n}, \ldots, \psi_0, \ldots, \psi_n\}$$

with $\psi_j(t) = e^{ijt}/\sqrt{2\pi}$; and let P_n denote the orthogonal projection of $L^2(0, 2\pi)$ onto V_n (see Section 3.7.1). For $\varphi = \sum a_m\psi_m$, it is straightforward that

$$P_n\varphi(s) = \sum_{m=-n}^{n} a_m\psi_m(s),$$

the truncation of the Fourier series for φ.

Recall the decomposition (13.3.4)–(13.3.5) of (13.3.1),

$$\mathcal{A}\varphi + B\varphi = g \tag{13.3.8}$$

with $\mathcal{A}\varphi$ given in (13.2.32). It is immediate that

$$P_n\mathcal{A} = \mathcal{A}P_n, \qquad P_n\mathcal{A}^{-1} = \mathcal{A}^{-1}P_n. \tag{13.3.9}$$

Approximate (13.3.8) by the equation

$$P_n(\mathcal{A}\varphi_n + B\varphi_n) = P_n g, \quad \varphi_n \in V_n. \tag{13.3.10}$$

Letting

$$\varphi_n(s) = \sum_{m=-n}^{n} a_m^{(n)} \psi_m(s)$$

and recalling (13.2.34), the equation (13.3.10) implies that the coefficients $\{a_m^{(n)}\}$ are determined from the linear system

$$\begin{aligned} \frac{a_k^{(n)}}{\max\{1,|k|\}} + \sum_{m=-n}^{n} a_m^{(n)} \int_0^{2\pi} \int_0^{2\pi} b(t,s)\psi_m(s)\overline{\psi_k(t)}\, ds\, dt \\ = \int_0^{2\pi} g(t)\overline{\psi_k(t)}\, dt, \quad k = -n, \ldots, n. \end{aligned} \tag{13.3.11}$$

Generally these integrals must be evaluated numerically.

The equation (13.3.8) is equivalent to

$$\varphi + \mathcal{A}^{-1}B\varphi = \mathcal{A}^{-1}g. \tag{13.3.12}$$

The right side function $\mathcal{A}^{-1}g \in L^2(0, 2\pi)$, by (13.3.6) and by the earlier assumption that $g \in H^1(2\pi)$. From the discussion following (13.3.7), $\mathcal{A}^{-1}B$ is a compact mapping from $L^2(0, 2\pi)$ into $L^2(0, 2\pi)$, and thus (13.3.12) is a Fredholm integral equation of the second kind. By the earlier assumption on the unique solvability of (13.3.8), we have $(I+\mathcal{A}^{-1}B)^{-1}$ exists on $L^2(0, 2\pi)$ to $L^2(0, 2\pi)$.

Using (13.3.9), the approximating equation (13.3.10) is equivalent to

$$\varphi_n + P_n\mathcal{A}^{-1}B\varphi_n = P_n\mathcal{A}^{-1}g. \tag{13.3.13}$$

Equation (13.3.13) is simply a standard Galerkin method for solving the equation (13.3.12), and it is exactly of the type discussed in Subsection 12.2.4.

Since $P_n\varphi \to \varphi$, for all $\varphi \in L^2(0, 2\pi)$, and since $\mathcal{A}^{-1}B$ is a compact operator, we have

$$\left\|(I - P_n)\mathcal{A}^{-1}B\right\| \to 0 \quad \text{as } n \to \infty$$

from Lemma 12.1.4 in Subsection 12.1.3 of Chapter 12. Then by standard arguments, the existence of $(I+\mathcal{A}^{-1}B)^{-1}$ implies that of $(I+P_n\mathcal{A}^{-1}B)^{-1}$, for all sufficiently large n. This is simply a repetition of the general argument given in Theorem 12.1.2, in Subsection 12.1.3. From (12.1.24) of that theorem,

$$\|\varphi - \varphi_n\|_0 \le \left\|(I + P_n\mathcal{A}^{-1}B)^{-1}\right\| \|\varphi - P_n\varphi\|_0, \tag{13.3.14}$$

where $\|\cdot\|_0$ is the norm for $H^0(2\pi) \equiv L^2(0, 2\pi)$. For more detailed bounds on the rate of convergence, apply Theorem 7.5.7 of Section 7.5, obtaining

$$\|\varphi - \varphi_n\|_0 \leq \frac{c}{n^q}\|\varphi\|_q, \quad \varphi \in H^q(2\pi) \tag{13.3.15}$$

for any $q > 0$.

A fully discretized variant of (13.3.13) is given in [18, p. 351], including numerical examples.

Exercise 13.3.1 Let k be a non-negative integer. Solve the integral equation

$$-\frac{1}{\pi}\int_0^{2\pi} \varphi(s) \log\left|2e^{-1/2}\sin\left(\frac{t-s}{2}\right)\right| ds = \cos(kt), \quad 0 \leq t \leq 2\pi.$$

Exercise 13.3.2 Obtain an explicit formula for the function $b(t, s)$ when the boundary S is the ellipse of (13.2.36). Simplify it as much as possible.

Suggestion for Further Reading.

Parts of this chapter are modifications of portions of Atkinson [18, Chap. 7]. Chapters 7–9 of the latter contain a more complete and extensive introduction to boundary integral equation reformulations and their numerical solution, again for only Laplace's equation; and a very large set of references are given there. More complete introductions to boundary integral equations and their analysis can be found in Kress [149], Mikhlin [172], and Pogorzelski [187]. From the perspective of applications of BIE, see Jaswon and Symm [130] and Pozrikidis [188].

A comprehensive survey of numerical methods for planar BIE of both the first and second kinds is given by Sloan [206]. An important approach to the study and solution of BIE, one which we have omitted here, is to regard BIEs as strongly elliptic pseudo-differential operator equations between suitably chosen Sobolev spaces. Doing such, we can apply Galerkin and finite-element methods to the BIE, in much the manner of Chapters 9 and 10. There is no other numerical method known for solving and analyzing some BIEs. As important examples of this work, see Wendland [229], [230], [231]. An introduction is given in [18, Section 7.4].

14

Multivariable Polynomial Approximations

In Chapter 3 we introduced the approximation of univariate functions by polynomials and trigonometric functions (see Sections 3.4–3.7). In this chapter we extend those ideas to multivariable functions and multivariable polynomials. In the univariate case there are only three types of approximation domains, namely $[a, b]$, $[a, \infty)$, and $(-\infty, \infty)$, with a and b finite. In contrast, there are many types of approximation domains in the multivariable case. This chapter is only an introduction to this area, and to make it more accessible, we emphasize planar problems. In particular, we consider the unit disk

$$\mathbb{B}_2 = \left\{(x, y) \mid x^2 + y^2 \leq 1\right\}$$

as the principal approximation domain of interest.

14.1 Notation and best approximation results

For notation, we use (x, y) to specify points when dealing with planar problems; but when talking about functions defined in $\mathbb{R}^d$, with some $d \geq 2$, we use a column vector to denote a point $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x} = (x_1, \ldots, x_d)^T$. We recall notation from Chapter 1. For a closed set $D \subset \mathbb{R}^d$ with non-empty interior, the space $C(D)$ consists of all functions f that are continuous on D; the norm is $\|\cdot\|_\infty$. Similarly, $L^2(D)$ is the set of functions that are square integrable over D with the standard inner product topology. For

K. Atkinson and W. Han, *Theoretical Numerical Analysis: A Functional Analysis Framework*, Texts in Applied Mathematics 39, DOI: 10.1007/978-1-4419-0458-4_14,
© Springer Science + Business Media, LLC 2009

simplicity in notation,

$$\int_D f(\boldsymbol{x})\,dx \equiv \int_D f(x_1,\ldots,x_d)\,dx_1\cdots dx_d.$$

The space $C^k(D)$, $k \geq 1$, consists of functions which are k-times continuously differentiable over D. When D is lower dimensional, e.g. a surface or path in $\mathbb{R}^d$, then additional care must be taken with these definitions; e.g. see Subsection 7.2.3.

When dealing with polynomials, we use multi-index notation. Let $\alpha = (\alpha_1,\ldots,\alpha_d)$ with all $\alpha_j \geq 0$ being integers; and let $|\alpha| = \alpha_1 + \cdots + \alpha_d$. The monomial $\boldsymbol{x}^\alpha$ is defined by

$$\boldsymbol{x}^\alpha = x_1^{\alpha_1}\cdots x_d^{\alpha_d}.$$

Introduce

$$\Pi_n^d = \left\{ \sum_{|\alpha|\leq n} b_\alpha \boldsymbol{x}^\alpha \;\middle|\; b_\alpha \in \mathbb{R} \right\}.$$

These are the polynomials in d variables that are of degree $\leq n$. For $d = 2$, it is straightforward to show

$$\dim \Pi_n^2 \equiv \left|\Pi_n^2\right| = \frac{1}{2}\,(n+1)\,(n+2)\,.$$

In general,

$$\left|\Pi_n^d\right| = \binom{n+d}{d}.$$

Assume $D \subset \mathbb{R}^d$ is a closed and bounded set. For a function $f \in C(D)$, the minimax approximation error when using Π_n^d as an approximating space is defined by

$$E_n(f) = \inf_{p\in\Pi_n^d} \|f-p\|_\infty. \tag{14.1.1}$$

The work of generalizing Jackson's Theorem 3.7.2 to multivariable polynomial approximation has a long history, beginning with the paper of Gronwall [100] in 1914. We give results from Ragozin [190]. As notation when using partial derivatives, we write

$$\begin{aligned}
\partial^\alpha f &= \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d}}, \quad \alpha = (\alpha_1,\ldots\alpha_d)\,,\\
\|f\|_{*,n} &= \max_{|\alpha|\leq n} \|\partial^\alpha f\|_\infty,\\
\omega(f,h) &= \sup_{|\boldsymbol{x}-\boldsymbol{y}|\leq h} |f(\boldsymbol{x}) - f(\boldsymbol{y})|\,,\\
\omega_n(f,h) &= \sum_{|\alpha|=n} \omega(\partial^\alpha f, h).
\end{aligned}$$

Also, introduce the unit ball,

$$\mathbb{B}_d = \left\{ \boldsymbol{x} \in \mathbb{R}^d \mid x_1^2 + \cdots + x_d^2 \leq 1 \right\}$$

with $d \geq 2$.

Theorem 14.1.1 *Assume $f \in C^k(\mathbb{B}_d)$. Then there exist polynomials $p_n \in \Pi_n^d$ for which*

$$\|f - p_n\|_\infty \leq \frac{c(k,d)}{n^k} \left[\frac{\|f\|_{*,k}}{n} + \omega_k\left(f, \frac{1}{n}\right) \right]. \tag{14.1.2}$$

The constant $c(k,d)$ depends on only k and d.

This is taken from [190, Thm. 3.4]. The proof is quite complicated and we only reference it. The theorem generalizes (3.7.4) of Jackson's Theorem 3.7.2 for univariate polynomial approximation. Another approach to this problem uses a connection between approximations of functions on $\mathbb{S}^d$ and those on $\mathbb{B}_d$. For a complete development of this approach, including the measuring of error using norms other than $\|\cdot\|_\infty$, see Rustamov [201] and Yuan Xu [240], [241].

Exercise 14.1.1 Show that the results of this section apply also to polynomial approximations over the ellipse

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 \leq 1,$$

where $a, b > 0$.

14.2 Orthogonal polynomials

In this section, we follow closely the ideas of Yuan Xu [239]. Consider Π_n^d as a subspace of $L^2(\mathbb{B}_d)$. Introduce

$$\mathbb{P}_n^d = \operatorname{span}\left\{ \boldsymbol{x}^\alpha \mid \alpha = (\alpha_1, \ldots \alpha_d),\ |\alpha| = n \right\}, \quad n \geq 0,$$

the homogeneous polynomials of degree n. Easily,

$$\Pi_n^d = \mathbb{P}_0^d \oplus \mathbb{P}_1^d \oplus \cdots \oplus \mathbb{P}_n^d. \tag{14.2.1}$$

Also, introduce

$$\mathbb{V}_n^d = \left\{ p \in \Pi_n^d \mid (p,q) = 0 \ \forall\, q \in \Pi_{n-1}^d \right\}$$

for $n \geq 1$; and define $\mathbb{V}_0^d = \{c \mid c \in \mathbb{R}\}$.

Lemma 14.2.1

$$\Pi_n^d = \mathbb{V}_0^d \oplus \mathbb{V}_1^d \oplus \cdots \oplus \mathbb{V}_n^d. \tag{14.2.2}$$

Proof. The proof is by induction on n. Clearly it is true for $n = 0$. Assume it is true for $n = k - 1$ and prove it for $n = k$.

For notation, let $N_n = \dim \Pi_n^d$, $Q_n = \dim \mathbb{P}_n^d$, and $M_n = \dim \mathbb{V}_n^d$, $n \geq 0$. It is immediate from (14.2.1) that

$$N_k = N_{k-1} + Q_k. \tag{14.2.3}$$

From the fact that all elements of $\mathbb{V}_k^d$ are polynomials of degree k, it follows that

$$\Pi_{k-1}^d \oplus \mathbb{V}_k^d \subset \Pi_k^d, \tag{14.2.4}$$

and this implies

$$N_{k-1} + M_k \leq N_k.$$

When combined with (14.2.3), this implies $M_k \leq Q_k$. We want to show $M_k = Q_k$. Once having done so, it follows that $\dim \Pi_k^d = \dim \Pi_{k-1}^d + \dim \mathbb{V}_k^d$. When combined with (14.2.4), this implies $\Pi_{k-1}^d \oplus \mathbb{V}_k^d = \Pi_k^d$, completing the induction.

For each basis element $\boldsymbol{x}^\alpha \in \mathbb{P}_k^d$, $\alpha = (\alpha_1, \ldots, \alpha_d)$, $|\alpha| = k$, consider constructing a polynomial

$$p(x) = \boldsymbol{x}^\alpha + r(\boldsymbol{x}), \quad r \in \Pi_{k-1}^d = \mathbb{V}_0^d \oplus \cdots \oplus \mathbb{V}_{k-1}^d$$

with the requirement that p be orthogonal to every element of Π_{k-1}^d. Each $\mathbb{V}_m^d$ can have an orthogonal basis $\{\varphi_{m,j} \mid j = 1, \ldots, M_m\}$; and by means of the induction hypothesis, these can be used collectively to form an orthogonal basis for Π_{k-1}^d. The polynomial r can be written in terms of this basis, say

$$r = \sum_{m=0}^{k-1} \sum_{j=1}^{M_\ell} \alpha_{m,j} \varphi_{m,j}.$$

When we impose the requirement that p be orthogonal to each element of our orthogonal basis for Π_{k-1}^d, the polynomial r is determined uniquely:

$$\alpha_{m,j} = -\frac{(\boldsymbol{x}^\alpha, \varphi_{m,j})}{(\varphi_{m,j}, \varphi_{m,j})}, \quad j = 1, \ldots, M_{k-1}, \quad m = 0, \ldots, k-1.$$

This establishes a one-to-one correspondence between the polynomial space $\mathbb{P}_k^d = \text{span}\,\{\boldsymbol{x}^\alpha \mid |\alpha| = k\}$ and a subspace of $\mathbb{V}_k^d$, implying $Q_k \leq M_k$. When combined with the earlier and reverse inequality, we have $M_k = Q_k$. From the earlier argument, it then follows that $\Pi_k^d = \Pi_{k-1}^d \oplus \mathbb{V}_k^d$, thus completing the induction. □

The decomposition (14.2.2) allows the creation of an orthogonal basis for Π_n^d by first constructing an orthogonal basis for each $\mathbb{V}_n^d$. In the univariate

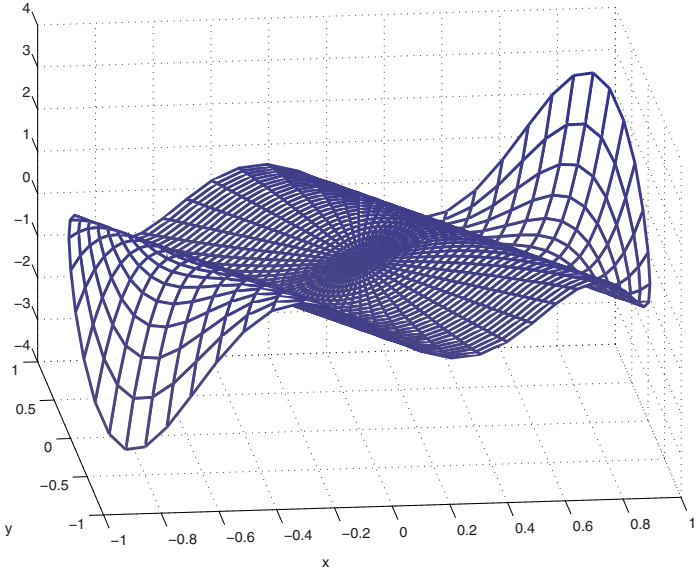

FIGURE 14.1. The orthonormal *ridge polynomial* $\Phi_{5,1}(x, y)$

case, each subspace $\mathbb{V}_n^1$ has dimension 1, and therefore each $\mathbb{V}_n^1$ has a unique basis (up to multiplication by a nonzero constant). In the multivariable case,

$$\dim \mathbb{V}_n^d = \binom{n+d-1}{d-1},$$

which is greater than 1 when $d > 1$. For the planar case, $d = 2$ and $\dim \mathbb{V}_n^d = n + 1$. There are many choices of an orthogonal basis for the subspace $\mathbb{V}_n^d$ when $d > 1$, and several are listed in [239, Section 1.2].

Example 14.2.2 Define the polynomials

$$\varphi_{n,k}(x, y) = \frac{1}{\sqrt{\pi}} U_n(x \cos(kh) + y \sin(kh)), \quad (x, y) \in \mathbb{B}_2, \quad h = \frac{\pi}{n+1} \tag{14.2.5}$$

for $k = 0, 1, \ldots, n$. The function U_n is the Chebyshev polynomial of the second kind of degree n:

$$U_n(t) = \frac{\sin(n+1)\,\theta}{\sin\theta}, \quad t = \cos\theta, \quad -1 \le t \le 1, \quad n = 0, 1, \ldots.$$

This set of polynomials form an orthonormal basis for $\mathbb{V}_n^2$ with respect to the standard inner product over $\mathbb{B}_2$. For a proof, see [160]. Using (14.2.2), the set $\{\varphi_{n,k} \mid 0 \le k \le n \le N\}$ is an orthonormal basis of $\Pi_N^2 \subset L^2(\mathbb{B}_2)$.

These polynomials are sometimes referred to as "ridge polynomials". Figure 14.1 gives a graphic of $\varphi_{5,1}(x, y)$. It also illustrates the reason for the name given to these polynomials, that their graphs contain parallel ridges, with $\varphi_{n,k}(x, y)$ constant along the line

$$x\cos(kh) + y\sin(kh) = t \tag{14.2.6}$$

with t constant. □

14.2.1 Triple recursion relation

The triple recursion relation for univariate orthogonal polynomials was developed in Problems 3.5.5–3.5.6 of Section 3.5. We consider here the generalization to multivariable orthogonal polynomials. The generalization gives a relationship between orthogonal bases for $\mathbb{V}_{n-1}^d$, $\mathbb{V}_n^d$, and $\mathbb{V}_{n+1}^d$. To simplify the presentation, we work with only the planar case; but the results generalize in a straightforward way to higher dimensions.

Let $\mathbb{V}_n^2$ have the orthogonal basis $\{\varphi_{n,k} \mid k = 0, 1, \ldots, n\}$, $n \geq 0$, and it need not be the basis given in (14.2.5). Introduce the vector function

$$\boldsymbol{p}_n(\boldsymbol{x}) = (\varphi_{n,0}(\boldsymbol{x}), \varphi_{n,1}(\boldsymbol{x}), \ldots, \varphi_{n,n}(\boldsymbol{x}))^T.$$

In the following theorem, we replace the planar point (x, y) with $\boldsymbol{x} = (x_1, x_2)$. Also, let $\boldsymbol{p}_{-1}(\boldsymbol{x}) = 0$.

Theorem 14.2.3 *Let $\mathbb{V}_n^2$ have the orthogonal basis $\{\varphi_{n,k} \mid k = 0, 1, \ldots, n\}$, $n \geq 0$. For $n \geq 0$ and for $j = 1, 2$, there exist unique matrices $A_{n,j}$, $B_{n,j}$ and $C_{n,j}$ with the respective orders $(n+1)\times(n+2)$, $(n+1)\times(n+1)$, and $(n+1)\times n$, for which*

$$x_j\boldsymbol{p}_n(\boldsymbol{x}) = A_{n,j}\boldsymbol{p}_{n+1}(\boldsymbol{x}) + B_{n,j}\boldsymbol{p}_n(\boldsymbol{x}) + C_{n,j}\boldsymbol{p}_{n-1}(\boldsymbol{x}), \quad j = 1, 2 \tag{14.2.7}$$

with $C_{-1,j} = 0$. If the polynomials $\{\varphi_{n,k} \mid k = 0, 1, \ldots, n\}$, $n \geq 0$, are orthonormal, then $C_{n,j} = A_{n-1,j}^T$.

Proof. Consider the polynomial components of $x_j\boldsymbol{p}_n(\boldsymbol{x})$, namely $x_j\varphi_{n,k}(\boldsymbol{x})$ for some k, $0 \leq k \leq n$. Each such component is a polynomial of degree $n+1$. As such, we can write it as a combination of the orthogonal basis polynomials from $\mathbb{V}_0^2$, $\mathbb{V}_1^2$, …, $\mathbb{V}_{n+1}^2$:

$$x_j\varphi_{n,k}(\boldsymbol{x}) = \sum_{m=0}^{n+1}\sum_{\ell=0}^{m}\alpha_{m,\ell}\varphi_{m,\ell}(\boldsymbol{x}).$$

Then

$$\alpha_{m,\ell} = \int_{\mathbb{B}_2} x_j\varphi_{n,k}(\boldsymbol{x})\varphi_{m,\ell}(\boldsymbol{x})\,dx \Big/ \int_{\mathbb{B}_2}\varphi_{m,\ell}(\boldsymbol{x})^2 dx.$$

Note that $x_j\varphi_{m,\ell}(\boldsymbol{x})$ has degree $m+1$. As such, $\varphi_{n,k}(\boldsymbol{x})$ is orthogonal to it when $m+1<n$; and thus $\alpha_{m,\ell}=0$ when $m<n-1$. As a consequence, $x_j\varphi_{n,k}(\boldsymbol{x})$ is a combination of basis functions from $\mathbb{V}_{n-1}^d$, $\mathbb{V}_n^d$, and $\mathbb{V}_{n+1}^d$. This yields (14.2.7).

We can obtain formulas for $A_{n,j}$, $B_{n,j}$ and $C_{n,j}$ as well. We begin by considering the matrix H_n of order $n\times n$,

$$H_n=\int_{\mathbb{B}_2}\boldsymbol{p}_n(\boldsymbol{x})\,\boldsymbol{p}_n(\boldsymbol{x})^T\,dx,$$
$$(H_n)_{k,\ell}=\int_{\mathbb{B}_2}\varphi_{n,k}(\boldsymbol{x})\,\varphi_{n,\ell}(\boldsymbol{x})\,dx,\quad 0\le k,\ell\le n.$$

This is a Gram matrix of orthogonal polynomials, and thus it is nonsingular and diagonal. If $\{\varphi_{n,k}\mid k=0,1,\ldots,n\}$ is an orthonormal basis, then $H_n=I$. The matrices $A_{n,j}$, $B_{n,j}$ and $C_{n,j}$ can be obtained by solving

$$A_{n,j}H_{n+1}=\int_{\mathbb{B}_2}x_j\boldsymbol{p}_n(\boldsymbol{x})\,\boldsymbol{p}_{n+1}(\boldsymbol{x})^T\,dx, \tag{14.2.8}$$
$$B_{n,j}H_n=\int_{\mathbb{B}_2}x_j\boldsymbol{p}_n(\boldsymbol{x})\,\boldsymbol{p}_n(\boldsymbol{x})^T\,dx, \tag{14.2.9}$$
$$A_{n,j}H_{n+1}=H_nC_{n+1,j}^T. \tag{14.2.10}$$

The proofs are left as an exercise for the reader. When the polynomials $\{\varphi_{n,k}\mid k=0,1,\ldots,n\}$, $n\ge 0$, are orthonormal, it follows from (14.2.10) that $C_{n,j}=A_{n-1,j}^T$. □

As with univariate orthonormal polynomials in Subsection 3.7.2, we can generalize to multivariable orthogonal polynomials the idea of reproducing kernel and the Christoffel–Darboux identity; see [239, p. 146].

The preceding material also can be generalized to a weighted inner product

$$(f,g)_w=\int_{\mathbb{B}_2}w(\boldsymbol{x})\,f(\boldsymbol{x})\,g(\boldsymbol{x})\,dx. \tag{14.2.11}$$

for a weight functions $w(x)$ satisfying standard assumptions. Two other important domains are the unit simplex,

$$\mathbb{T}_2=\{(x,y)\mid x,y\ge 0,\,x+y\le 1\},$$

and the unit sphere in $\mathbb{R}^3$,

$$\mathbb{S}^2=\left\{(x,y,z)\mid x^2+y^2+z^2=1\right\},$$

and their extensions to higher dimensions. There are close connections between orthogonal polynomials on $\mathbb{B}_2$ and those on $\mathbb{T}_2$ and $\mathbb{S}^2$; and the use of the weighted inner products (14.2.11) is necessary in developing those connections. For these and many other results on multivariable orthogonal polynomials, see the lecture notes of Yuan Xu [239] and the book of Dunkl and Xu [72].

14.2.2 The orthogonal projection operator and its error

As earlier, let $\{\varphi_{m,k} \mid k = 0, 1, \ldots, m,\ m = 0, \ldots, n\}$ be an orthonormal basis for Π_n^2, $n \geq 0$. The orthogonal projection operator P_n from $L^2(\mathbb{B}_2)$ onto Π_n^2 is given by

$$P_n f(\boldsymbol{x}) = \sum_{m=0}^{n} \sum_{\ell=0}^{m} (f, \varphi_{m,\ell})\, \varphi_{m,\ell}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}_2,\ f \in L^2(\mathbb{B}_2). \tag{14.2.12}$$

This converges in the inner product norm of $L^2(\mathbb{B}_2)$ because of the completeness of the polynomials in the space $L^2(\mathbb{B}_2)$. For any $q \in \Pi_n^2$, we have

$$f - P_n f = f - q - P_n(f - q).$$

Use $\|P_n\|_{L^2 \to L^2} = 1$ to obtain

$$\|f - P_n f\|_{L^2} \leq 2\|f - q\|_{L^2} \leq 2\pi \|f - q\|_\infty,$$

provided $f \in C(\mathbb{B}_2)$. Combine this with the defnition in (14.1.1) to obtain

$$\|f - P_n f\|_{L^2} \leq 2\pi E_n(f), \quad f \in C(\mathbb{B}_2).$$

This can be combined with the results of Theorem 14.1.1 to obtain rates of convergence for the L^2 error in approximating f by $P_n f$.

In addition, we would like to know when $P_n f$ converges uniformly to f, i.e. in the norm of $C(\mathbb{B}_2)$. As for univariate approximations, given in (3.7.15) of Subsection 3.7.2, we can derive

$$\|f - P_n f\|_\infty \leq (1 + \|P_n\|)\, E_n(f), \quad f \in C(\mathbb{B}_2), \tag{14.2.13}$$

where P_n is considered as an operator from $C(\mathbb{B}_2)$ onto $\Pi_n^2 \subset C(\mathbb{B}_2)$. Thus we need to bound $\|P_n\|_{C \to C}$ in order to bound the rate of uniform convergence of $P_n f$ to f.

From [238], the function $\mathcal{P}_n f$ can be written as an integral operator,

$$P_n f(\boldsymbol{x}) = \int_{\mathbb{B}_2} G_n(\boldsymbol{x}, \boldsymbol{y})\, f(\boldsymbol{y})\, dy, \quad \boldsymbol{x} \in \mathbb{B}_2. \tag{14.2.14}$$

The function G_n is the reproducing kernel for Π_n^2 and is given by

$$G_n(\boldsymbol{x}, \boldsymbol{y}) = c_n \int_0^\pi P_n^{(3/2,1/2)}\left(\boldsymbol{x} \cdot \boldsymbol{y} + \sqrt{1 - \|\boldsymbol{x}\|^2}\sqrt{1 - \|\boldsymbol{y}\|^2}\cos\psi\right) d\psi. \tag{14.2.15}$$

where

$$c_n = \frac{\sqrt{\pi}\,\Gamma(n+3)}{4\,\Gamma(n+3/2)}.$$

The function $P_n^{(3/2,1/2)}(t)$ is the Jacobi polynomial of degree n associated with the weight function $(1-t)^{3/2}(1+t)^{1/2}$; see (3.5.3) in Section 3.5.

Theorem 14.2.4 *When considering the orthogonal projection operator P_n as an operator from $C(\mathbb{B}_2)$ to $\Pi_n^2 \subset C(\mathbb{B}_2)$,*

$$\|P_n\| = \mathcal{O}(n). \tag{14.2.16}$$

From the earlier result (2.2.8) in Section 2.2, the norm of P_n when regarded as an integral operator from $C(\mathbb{B}_2)$ to $C(\mathbb{B}_2)$ is given by

$$\|P_n\| = \sup_{\boldsymbol{x}\in\mathbb{B}_2} \int_{\mathbb{B}_2} |G_n(\boldsymbol{x},\boldsymbol{y})|\, dy.$$

The proof of both (14.2.15) and (14.2.16) is a special case of more general results given in [238]. When combined with (14.2.13) and Theorem 14.1.1, we have the following result.

Theorem 14.2.5 *Assume $f \in C^k(\mathbb{B}_2)$ with $k \geq 1$. Then*

$$\|f - P_n f\|_\infty \leq \frac{c_1(k)}{n^{k-1}} \left[\frac{\|f\|_{*,k}}{n} + \omega_k\left(f, \frac{1}{n}\right)\right]. \tag{14.2.17}$$

The constant $c_1(k)$ depends on only k. If $f \in C^1(\mathbb{B}_2)$, then $P_n f$ converges uniformly to f.

Combining the result (14.2.16) with the Banach-Steinhaus theorem, Theorem 2.4.5, implies there is at least one function $f \in C(\mathbb{B}_2)$ for which the sequence of functions $P_n f$ does not converge uniformly to f; it will converge, however, in the norm of $L^2(\mathbb{B}_2)$.

Exercise 14.2.1 Prove the formulas (14.2.8)–(14.2.10).

Exercise 14.2.2 Let $D = \mathbb{B}_2$.
(a). Show that $\{1, x, y\}$ is an orthogonal basis for Π_n^2.
(b). The set $\{x^2, xy, y^2\}$ is a basis of $\mathbb{P}_2^2$. Using the Gram-Schmidt method in the proof of Theorem 1.3.16 in Chapter 1, construct an orthonormal basis for $\mathbb{V}_2^2$.

Exercise 14.2.3 Using (14.2.5), construct explicitly an orthonormal basis for $\mathbb{V}_2^2$. Show directly that it is orthogonal to $\mathbb{V}_1^2$ and $\mathbb{V}_0^2$.

Exercise 14.2.4 Let $f(x, y, z)$ denote an arbitrary function that is integrable over $\mathbb{S}^2$. Show

$$\int_{\mathbb{S}^2} f(x,y,z)\, dS = \int_{\mathbb{B}_2} \frac{f\left(x, y, \sqrt{1-x^2-y^2}\right) + f\left(x, y, -\sqrt{1-x^2-y^2}\right)}{\sqrt{1-x^2-y^2}}\, dx\, dy.$$

This identity can be used to connect orthonormal spherical harmonics (see the discussion of spherical polynomials and spherical harmonics given in Section 7.5.5

of Chapter 7) to corresponding orthonormal bases of Π_n^2 for the weighted inner products

$$(g,h)_1 = \int_{\mathbb{B}_2} \frac{g(x,y)\,h(x,y)}{\sqrt{1-x^2-y^2}}\,dx\,dy,$$
$$(g,h)_2 = \int_{\mathbb{B}_2} g(x,y)\,h(x,y)\,\sqrt{1-x^2-y^2}\,dx\,dy.$$

For details of this relationship, see Xu [237].

14.3 Hyperinterpolation

In practice the formula (14.2.12) for $P_n f(\boldsymbol{x})$ must be evaluated by using numerical integration to approximate the coefficients $(f, \varphi_{m,\ell})$. The resulting approximation of $P_n f$ is sometimes referred to as a "discrete orthogonal projection". It has also become of interest for another reason. In one variable there is a well-developed theory of polynomial interpolation, and how to choose the node points is well-understood. In contrast, very little is understood about how best to carry out interpolation over $\mathbb{B}_2$ or over the unit sphere $\mathbb{S}^2$. For that reason, we use a quadrature based approximation of $P_n f(\boldsymbol{x})$, and this process is commonly referred to as "hyperinterpolation". It is best to choose a quadrature formula which satisfies certain properties. We will give one such quadrature method, and we will point out some of the resulting special properties.

For our quadrature over the unit disk, we use

$$\int_{\mathbb{B}_2} f(\boldsymbol{x})\,dx \approx \frac{2\pi}{2n+1} \sum_{l=0}^{n} \sum_{m=0}^{2n} \omega_l r_l f\left(r_l, \frac{2\pi\,m}{2n+1}\right). \tag{14.3.1}$$

The trapezoidal rule is used for the integration in the azimuthal direction and a Gaussian quadrature rule is used for the radial direction. This quadrature is exact for all polynomials $f \in \Pi_{2n+1}^2$. Here the numbers ω_l are the weights of the $(n+1)$-point Gauss-Legendre quadrature on $[0,1]$, and thus

$$\int_0^1 p(x)\,dx = \sum_{l=0}^{n} p(r_l)\,\omega_l$$

for all single-variable polynomials $p(x)$ with $\deg(p) \le 2n+1$. Using (14.3.1) to approximate the inner product $(\cdot,\cdot)$ of $L^2(\mathbb{B}_2)$, introduce the discrete inner product

$$(f,g)_n = \frac{2\pi}{2n+1} \sum_{l=0}^{n} \sum_{m=0}^{2n} \omega_l r_l f\left(r_l, \frac{2\pi\,m}{2n+1}\right) g\left(r_l, \frac{2\pi\,m}{2n+1}\right). \tag{14.3.2}$$

The weights in this discrete inner product are all positive. Moreover,

$$(f,g)_n = (f,g) \quad \forall\, f,g \in \Pi_n^2. \tag{14.3.3}$$

With the help of the discrete inner product we can now define an approximation to the orthogonal projection $P_n f$ when f is restricted to being continuous over $\mathbb{B}_2$:

$$L_n f(\boldsymbol{x}) = \sum_{m=0}^{n} \sum_{\ell=0}^{m} (f, \varphi_{m,\ell})_n \, \varphi_{m,\ell}(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}_2,\ f \in C(\mathbb{B}_2). \tag{14.3.4}$$

From (14.3.3),

$$L_n f = f \quad \forall\, f \in \Pi_n^2. \tag{14.3.5}$$

The operator L_n is the *hyperinterpolation operator* of Sloan and Womersley [207]. It is also a projection operator from $C(\mathbb{B}_2)$ onto Π_n^2; it is an example of a *discrete orthogonal projection operator*, as in [20]. Galerkin methods using this approximating operator L_n are sometimes called *discrete Galerkin methods*.

14.3.1 The norm of the hyperinterpolation operator

By a method similar to that used for P_n in (14.2.14),

$$\|L_n\|_{C(D)\to C(D)} = \frac{2\pi}{2n+1} \max_{x\in\mathbb{B}_2} \sum_{l=0}^{n} \sum_{m=0}^{2n} \omega_l r_l \left| G_n(x, \xi_{l,m}) \right|. \tag{14.3.6}$$

The function G is the same as in (14.2.15). From this, it is shown in [117] that

$$\|L_n\|_{C(D)\to C(D)} = \mathcal{O}(n \log n).$$

This is very close to the rate of $\mathcal{O}(n)$ growth for $\|P_n\|_{C(D)\to C(D)}$, given in (14.2.16). As in (14.2.17), this leads to the rate of convergence result

$$\|f - L_n f\|_\infty \le \frac{c_2(k)\, \log n}{n^{k-1}} \left[\frac{\|f\|_{*,k}}{n} + \omega_k\left(f, \frac{1}{n}\right) \right], \quad n \ge 2. \tag{14.3.7}$$

Exercise 14.3.1 Prove (14.3.5).

14.4 A Galerkin method for elliptic equations

Consider solving the elliptic partial differential equation

$$Lu(\boldsymbol{x}) \equiv -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} \left[a_{i,j}(\boldsymbol{x}) \frac{\partial u(\boldsymbol{x})}{\partial x_j} \right] + \gamma(\boldsymbol{x})\, u(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}_2 \tag{14.4.1}$$

with the Dirichlet boundary condition

$$u(\boldsymbol{x}) \equiv 0, \quad \boldsymbol{x} \in \partial\mathbb{B}_2 = \mathbb{S}^1. \tag{14.4.2}$$

Assume the functions $\gamma(\boldsymbol{x})$, $f(\boldsymbol{x})$, $a_{i,j}(\boldsymbol{x})$ are several times continuously differentiable over $\mathbb{B}_2$. As usual, assume the matrix $A(\boldsymbol{x}) = (a_{i,j}(\boldsymbol{x}))$ is symmetric, and also assume it satisfies the strong ellipticity condition,

$$\boldsymbol{\xi}^T A(\boldsymbol{x})\,\boldsymbol{\xi} \geq c_0 \boldsymbol{\xi}^T \boldsymbol{\xi}, \quad \boldsymbol{x} \in \mathbb{B}_2,\ \boldsymbol{\xi} \in \mathbb{R}^2 \tag{14.4.3}$$

with $c_0 > 0$. Also assume $\gamma(\boldsymbol{x}) \geq 0$, $\boldsymbol{x} \in \mathbb{B}_2$.

The Dirichlet problem (14.4.1)–(14.4.2) has the following variational reformulation: Find $u \in H_0^1(\mathbb{B}_2)$ such that

$$\int_{\mathbb{B}_2} \left[\sum_{i,j=1}^{d} a_{i,j}(\boldsymbol{x}) \frac{\partial u(\boldsymbol{x})}{\partial x_j} \frac{\partial v(\boldsymbol{x})}{\partial x_i} + \gamma(\boldsymbol{x})\, u(\boldsymbol{x})\, v(\boldsymbol{x}) \right] dx$$
$$= \int_{\mathbb{B}_2} f(\boldsymbol{x})\, v(\boldsymbol{x})\, dx \quad \forall\, v \in H_0^1(\mathbb{B}_2). \tag{14.4.4}$$

Over the space $H_0^1(\mathbb{B}_2)$, introduce the bilinear form

$$\mathcal{A}(v,w) = \int_{\mathbb{B}_2} \left[\sum_{i,j=1}^{d} a_{i,j}(\boldsymbol{x}) \frac{\partial v(\boldsymbol{x})}{\partial x_j} \frac{\partial w(\boldsymbol{x})}{\partial x_i} + \gamma(\boldsymbol{x})\, v(\boldsymbol{x})\, w(\boldsymbol{x}) \right] dx \tag{14.4.5}$$

and the bounded linear functional

$$\ell(v) = \int_{\mathbb{B}_2} f(\boldsymbol{x})\, v(\boldsymbol{x})\, dx.$$

The variational problem (14.4.4) can now be written as follows: find $u \in H_0^1(\mathbb{B}_2)$ for which

$$\mathcal{A}(u,v) = \ell(v) \quad \forall\, v \in H_0^1(\mathbb{B}_2). \tag{14.4.6}$$

It is straightforward to show $\mathcal{A}$ is bounded,

$$|\mathcal{A}(v,w)| \leq c_{\mathcal{A}} \|v\|_1 \|w\|_1, \quad c_{\mathcal{A}} = \max_{\boldsymbol{x} \in \mathbb{B}_2} \|A(\boldsymbol{x})\|_2 + \|\gamma\|_\infty$$

with $\|\cdot\|_1$ the norm of $H_0^1(D)$ and $\|A(\boldsymbol{x})\|_2$ the matrix 2-norm of the matrix $A(\boldsymbol{x})$. In addition, we assume

$$\mathcal{A}(v,v) \geq c_e \|v\|_1^2, \quad v \in H_0^1(\mathbb{B}_2). \tag{14.4.7}$$

This follows generally from (14.4.3) and the size of the function $\gamma(\boldsymbol{x})$ over $\mathbb{B}_2$; when $\gamma \equiv 0$, $c_e = c_0$. Under standard assumptions on $\mathcal{A}$, including

the strong ellipticity in (14.4.7), the Lax-Milgram Theorem implies the existence of a unique solution u to (14.4.6) with

$$\|u\|_1 \leq \frac{1}{c_e}\|\ell\|.$$

Let $\mathbb{X}_n$ denote our approximation subspace,

$$\mathbb{X}_n = \left\{ \left(1 - x_1^2 - x_2^2\right) p(\boldsymbol{x}) \mid p \in \Pi_n^2 \right\}. \tag{14.4.8}$$

The subspaces Π_n and $\mathbb{X}_n$ have dimension

$$N_n = \frac{1}{2}(n+1)(n+2).$$

Lemma 14.4.1 *Let Δ denote the Laplacian operator in $\mathbb{R}^d$. Then*

$$\Delta : \mathbb{X}_n \xrightarrow[onto]{1-1} \Pi_n^2. \tag{14.4.9}$$

The proof of this is left as a problem.

14.4.1 The Galerkin method and its convergence

The Galerkin method for obtaining an approximate solution to (14.4.6) is as follows: find $u_n \in \mathbb{X}_n$ for which

$$\mathcal{A}(u_n, v) = \ell(v) \quad \forall\, v \in \mathbb{X}_n. \tag{14.4.10}$$

The Lax-Milgram Theorem (Section 8.3) implies the existence of u_n for all n. For the error in this Galerkin method, Cea's Lemma (Proposition 9.1.3) implies the convergence of u_n to u, and moreover,

$$\|u - u_n\|_1 \leq \frac{c_{\mathcal{A}}}{c_e} \inf_{v \in \mathbb{X}_n} \|u - v\|_1. \tag{14.4.11}$$

It remains to bound the best approximation error on the right side of this inequality.

Given an arbitrary $u \in H_0^2(\mathbb{B}_2)$, define $w = -\Delta u$. Then $w \in L^2(\mathbb{B}_2)$ and u satisfies the boundary value problem

$$\begin{aligned} -\Delta u(P) &= w(P), \quad P \in \mathbb{B}_2, \\ u(P) &= 0, \quad P \in \mathbb{S}^2. \end{aligned}$$

It follows that

$$u(P) = \int_{\mathbb{B}_2} G(P,Q)\, w(Q)\, dQ, \quad P \in \mathbb{B}_2. \tag{14.4.12}$$

The Green's function is defined by

$$G(P,Q) = \frac{1}{2\pi} \log \frac{|P-Q|}{|\mathcal{T}(P)-Q|}$$

as follows for $P \neq Q$, $Q \in \text{int}(\mathbb{B}_2)$, $P \in \mathbb{B}_2$. The quantity $\mathcal{T}(P)$ denotes the inverse point for P with respect to the unit sphere $\mathbb{S}^2$,

$$\mathcal{T}(r\boldsymbol{x}) = \frac{1}{r}\boldsymbol{x}, \quad 0 < r \leq 1, \ \boldsymbol{x} \in \mathbb{S}^2.$$

Differentiate (14.4.12) to obtain

$$\nabla u(P) = \int_{\mathbb{B}_2} [\nabla_P G(P,Q)]\, w(Q)\, dQ, \quad P \in \mathbb{B}_2.$$

Note that $\nabla_P G(P,\cdot)$ is absolutely integrable over $\mathbb{B}_2$, for all $P \in \mathbb{B}_2$.

Let $w_n \in \Pi_n^2$ be an approximation of w, say in the norm of either $C(\mathbb{B}_2)$ or $L^2(\mathbb{B}_2)$, and let

$$q_n(P) = \int_{\mathbb{B}_2} G(P,Q)\, w_n(Q)\, dQ, \quad P \in \mathbb{B}_2.$$

We can show $q_n \in \mathbb{X}_n$. This follows from Lemma 14.4.1 and noting that the mapping in (14.4.12) is the inverse of (14.4.9).

Then we have

$$u(P) - q_n(P) = \int_{\mathbb{B}_2} G(P,Q)\, [w(P) - w_n(Q)]\, dQ, \quad P \in \mathbb{B}_2,$$

$$\nabla\, [u(P) - q_n(P)] = \int_{\mathbb{B}_2} [\nabla_P G(P,Q)]\, [w(Q) - w_n(Q)]\, dQ, \quad P \in \mathbb{B}_2.$$

The integral operators on the right side are weakly singular compact integral operators on $L^2(\mathbb{B}_2)$ to $L^2(\mathbb{B}_2)$ [172, Chap. 7, Section 3]. This implies

$$\|u - q_n\|_1 \leq c\, \|w - w_n\|_0. \tag{14.4.13}$$

By letting w_n be the orthogonal projection of w into Π_n^2 (see (14.2.12)), the right side will go to zero since the polynomials are dense in $L^2(\mathbb{B}_2)$. In turn, this implies convergence in the $H_0^1(\mathbb{B}_2)$ norm for the right side in (14.4.11) provided $u \in H_0^2(\mathbb{B}_2)$.

The result

$$\inf_{v \in \mathbb{X}_n} \|u - v\|_1 \to 0 \quad \text{as } n \to \infty, \quad u \in H_0^2(\mathbb{B}_2)$$

can be extended to any $u \in H_0^1(\mathbb{B}_2)$. It basically follows from the denseness of $H_0^2(\mathbb{B}_2)$ in $H_0^1(\mathbb{B}_2)$. Let $u \in H_0^1(\mathbb{B}_2)$. We need to find a sequence of polynomials $\{q_n\}$ for which $\|u - q_n\|_1 \to 0$. We know $H_0^2(\mathbb{B}_2)$ is dense in

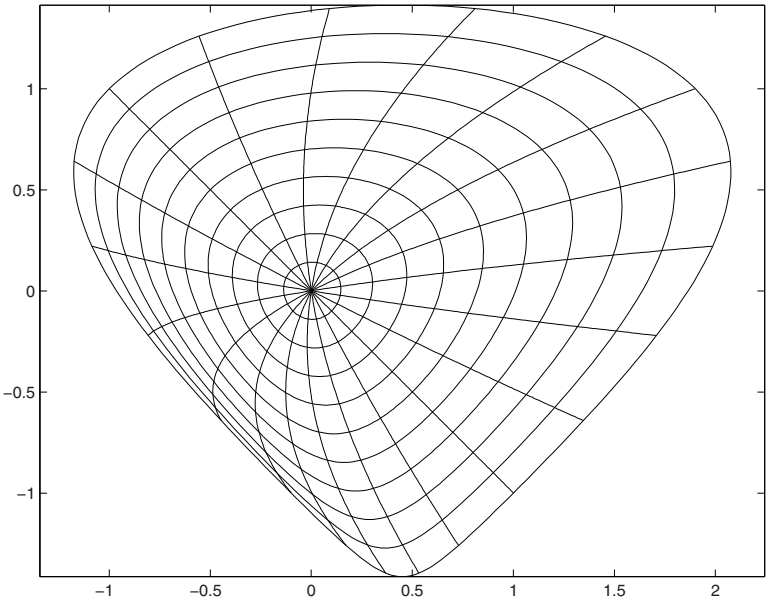

FIGURE 14.2. Images of (14.4.17) with $a = 0.5$, for lines of constant radius and constant azimuth on the unit disk

$H_0^1(\mathbb{B}_2)$. Given any $k > 0$, choose $u_k \in H_0^2(\mathbb{B}_2)$ with $\|u - u_k\|_1 \leq 1/k$. Then choose a polynomial w_k for which we have the corresponding polynomial q_k satisfying $\|u_k - q_k\|_1 \leq 1/k$, based on (14.4.13). (Regarding the earlier notation, q_k need not be of degree $\leq k$.) Then $\|u - q_k\|_1 \leq 2/k$. This completes the examination of the right hand side of (14.4.11) for any $u \in H_0^1(\mathbb{B}_2)$, and thus also, the bounding of the error in the Galerkin solution u_n.

Example 14.4.2 As a basis for Π_n^2, we choose the orthonormal ridge polynomials $\{\varphi_{m,k}(x, y) \mid 0 \leq k \leq m \leq n\}$ of (14.2.5), in which we revert to denoting a planar point by (x, y). As a basis for $\mathbb{X}_n$, use the functions

$$\psi_{m,k}(x, y) = \left(1 - x^2 - y^2\right) \varphi_{m,k}(x, y), \quad 0 \leq k \leq m \leq n. \tag{14.4.14}$$

As an example problem, we use a reformulation of the problem

$$-\Delta u(s, t) + e^{s-t} u(s, t) = g(s, t), \quad (s, t) \in D, \tag{14.4.15}$$

$$u(s, t) = 0, \quad (s, t) \in \partial D, \tag{14.4.16}$$

where D is the planar region determined by the mapping $\Phi : \mathbb{B}_2 \to D$, $(x, y) \mapsto (s, t)$,

$$s = x - y + ax^2, \quad t = x + y \tag{14.4.17}$$

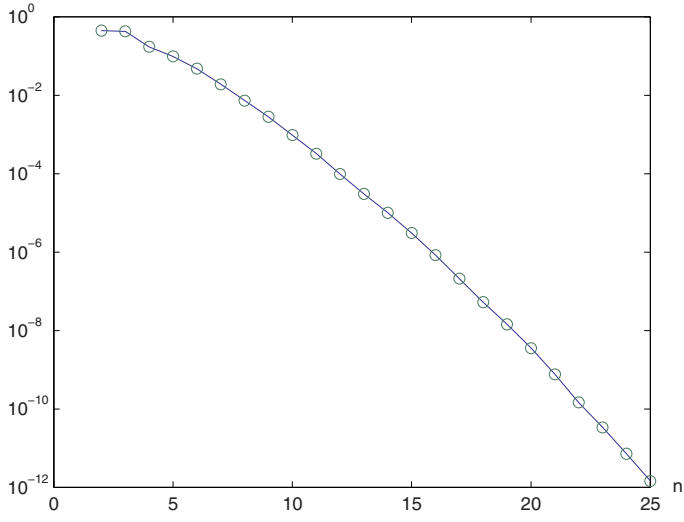

FIGURE 14.3. Error in Galerkin method (14.4.6) for (14.4.15)–(14.4.16)

with $0 < a < 1$ and $(x, y) \in \mathbb{B}_2$. The case $a = 0.5$ is pictured in Figure 14.2. The equation (14.4.15) converts to the elliptic equation (14.4.1) with

$$\begin{aligned} A(x,y) &= \frac{1}{1+ax}\begin{pmatrix} 1 & ax \\ ax & 2a^2x^2+2ax+1 \end{pmatrix}, \\ \gamma(x,y) &= 2\,(1+ax)\,e^{s-t}, \\ f(x,y) &= 2\,(1+ax)\,g(s,t), \end{aligned}$$

with (s, t) replaced using (14.4.17).

As a particular test case, we choose

$$u(s,t) = \left(1-x^2-y^2\right)\cos(\pi s), \quad (s,t) \in D \tag{14.4.18}$$

with (x, y) replaced using (14.4.17). The right-hand function g is obtained from u by applying (14.4.15). The errors in the Galerkin method (14.4.6) for solving (14.4.15)–(14.4.16) are given in Figure 14.3. The results exhibit an exponential rate of convergence.

For more information on the Galerkin method given in this section, including the use of a change of variables such as in (14.4.17), see [21].

Exercise 14.4.1 Prove Lemma 14.4.1.

Exercise 14.4.2 Let D be the ellipse $(x/a)^2 + (y/b)^2 \leq 1$. Consider solving the Dirichlet problem

$$-\Delta u(s,t) = f(s,t), \quad (s,t) \in \overset{\circ}{D},$$
$$u(s,t) = 0, \quad (s,t) \in \partial D.$$

Convert this to an equivalent problem over $\mathbb{B}_2$. Begin by introducing $v(x,y) = u(ax,by)$, $(x,y) \in \mathbb{B}_2$, and then find an elliptic equation for v over $\mathbb{B}_2$.

Exercise 14.4.3 Consider again the Dirichlet problem from Exercise 14.4.2, but for a general region D that is bounded and simply connected with a smooth boundary. Assume the existence of a mapping $\Phi : \mathbb{B}_2 \xrightarrow[onto]{1-1} D$, as was illustrated in (14.4.17); and assume that Φ is several times continuously differentiable and that

$$\det[J(x,y)] > 0, \quad (x,y) \in \mathbb{B}_2,$$

with $J(x,y)$ denoting the Jacobian matrix for Φ. Convert the Dirichlet problem over D to an equivalent elliptic problem over $\mathbb{B}_2$. Begin by introducing $v(x,y) = u(\Phi(x,y))$, $(x,y) \in \mathbb{B}_2$, and then find an elliptic equation for v over $\mathbb{B}_2$.

Suggestion for Further Readings. There is an extensive literature on univariate orthogonal polynomials. We note particularly the classic book of SZEGÖ [220] and the more recent book of ANDREWS, ASKEY AND ROY [4, Chaps. 5–7]. Although multivariable polynomial approximation theory has old roots, it has become much more popular in recent years. As noted earlier, there is a close connection between polynomial approximations on the unit sphere and the unit ball. This is developed in [241], applying results on the unit sphere to analogous results on the unit ball and unit simplex. The first results for best approximation on the unit sphere and for uniform convergence of the Laplace expansion appear to be those of GRONWALL [100]. A history of the subject is given in the landmark paper of RUSTAMOV [201]. An earlier important work for uniform polynomial approximations on both the unit sphere and the unit ball is that of RAGOZIN [190], [191]. A generalization of Ragozin's results to simultaneous approximation of a function and some of its derivatives is given in [31]. Orthogonal polynomials in two variables are discussed extensively by KOORNWINDER [145], for a variety of planar regions. For recent work on the theory of multivariable orthogonal polynomials, we recommend the survey paper of YUAN XU [239] and the book of DUNKL AND XU [72]. For another application of planar orthogonal polynomials, to the nonlinear Poisson equation, see [23].

References

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] M. Ainsworth and J. T. Oden, A Posteriori Error Estimation in Finite Element Analysis, John Wiley & Sons, Inc., New York, 2000.

[3] A. Aldroubi and M. Unser, eds., *Wavelets in Medicine and Biology*, CRC Press, Boca Raton, 1996.

[4] G. Andrews, R. Askey, and R. Roy, *Special Functions*, Cambridge University Press, Cambridge, 2001.

[5] P. Anselone, *Collectively Compact Operator Approximation Theory and Applications to Integral Equations*, Prentice-Hall, 1971.

[6] H. Anton and C. Rorres, *Elementary Linear Algebra*, 7^{th} ed., John Wiley, New York, 1994.

[7] T. Apostol, *Mathematical Analysis*, Addison-Wesley Pub, 1957.

[8] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, *SIAM J. Numer. Anal.* **39** (2002), 1749–1779.

[9] D. N. Arnold and W. Wendland, The convergence of spline collocation for strongly elliptic equations on curves, *Numer. Math.* **47** (1985), 317–341.

[10] K. Atkinson, The solution of non-unique linear integral equations, *Numer. Math.* **10** (1967), 117–124.

[11] K. Atkinson, The numerical solution of the eigenvalue problem for compact integral operators, *Trans. Amer. Math. Soc.* **129** (1967), 458–465.

[12] K. Atkinson, The numerical evaluation of fixed points for completely continuous operators, *SIAM J. Num. Anal.* **10** (1973), 799–807.

[13] K. Atkinson, Convergence rates for approximate eigenvalues of compact integral operators, *SIAM J. Num. Anal.* **12** (1975), 213–222.

[14] K. Atkinson, The numerical solution of a bifurcation problem, *SIAM J. Num. Anal.* **14** (1977), 584–599.

[15] K. Atkinson, *An Introduction to Numerical Analysis*, 2nd ed., John Wiley, New York, 1989.

[16] K. Atkinson, A survey of numerical methods for solving nonlinear integral equations, *J. Int. Eqns. & Applics* **4** (1992), 15–46.

[17] K. Atkinson, Two-grid iteration methods for linear integral equations of the second kind on piecewise smooth surfaces in $\mathbf{R}^3$, *SIAM J. Scientific Computing* **15** (1994), 1083–1104.

[18] K. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, UK, 1997.

[19] K. Atkinson, A personal perspective on the history of the numerical analysis of Fredholm integral equations of the second kind, in the proceedings of *The Birth of Numerical Analysis*, Leuven, Belgium, October 2007.

[20] K. Atkinson and A. Bogomolny, The discrete Galerkin method for integral equations, *Math. of Comp.* **48** (1987), 595–616 and S11–S15.

[21] K. Atkinson, D. Chien, and O. Hansen, A spectral method for elliptic equations: The Dirichlet problem, to appear in *Advances in Computational Mathematics*.

[22] K. Atkinson, I. Graham, and I. Sloan, Piecewise continuous collocation for integral equations, *SIAM J. Numer. Anal.* **20** (1983), 172–186.

[23] K. Atkinson and O. Hansen, Solving the nonlinear Poisson equation on the unit disk, *J. Integral Equations Appl.,* **17** (2005), 223–241.

[24] K. Atkinson and Y.-M. Jeon, Algorithm 788: Automatic boundary integral equation programs for the planar Laplace equation, *ACM Trans. on Math. Software* **24** (1998), 395–417.

[25] K. Atkinson and F. Potra, Projection and iterated projection methods for nonlinear integral equations, *SIAM J. Numer. Anal.* **24** (1987), 1352–1373.

[26] H. Attouch, G. Buttazzo, and G. Michaille, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*, SIAM and MPS, Philadelphia, 2006.

[27] J.-P. Aubin, *Applied Functional Analysis*, second edition, John Wiley, 1999.

[28] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1996.

[29] I. Babuška and A. K. Aziz, Survey lectures on the mathematical foundations of the finite element method, in A.K. Aziz, ed., *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972, 3–359.

[30] I. Babuška and T. Strouboulis, *The Finite Element Method and its Reliability*, Oxford University Press, 2001.

[31] T. Bagby, L. Bos, and N. Levenberg, Multivariate simultaneous approximation, *Constructive Approximation* **18** (2002), 569–577.

[32] J. W. Barrett and W.-B. Liu, Finite element approximation of the p-Laplacian, *Math. Comp.* **61** (1993), 523–537.

[33] M. Berger, *Nonlinearity and Functional Analysis*, Academic Press, 1977.

[34] J. Bergh and J. Löfström, *Interpolation Spaces, An Introduction*, Springer-Verlag, Berlin, 1976.

[35] C. Bernardi and Y. Maday, Spectral methods, in P. G. Ciarlet and J.-L. Lions, eds., *Handbook of Numerical Analysis*, Vol. V, North-Holland, Amsterdam, 1997, 209–485.

[36] M. Bernkopf, The development of function spaces with particular reference to their origins in integral equation theory, *Archive for History of Exact Sciences* **3** (1966), 1–96.

[37] G. Birkhoff, M. H. Schultz and R. S. Varga, Piecewise Hermite interpolation in one and two variables with applications to partial differential equations, *Numer. Math.* **11** (1968), 232–256.

[38] A. Boggess and F. J. Narcowich, *A First Course in Wavelets with Fourier Analysis*, Prentice Hall, NJ, 2001.

[39] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, third edition, Cambridge University Press, Cambridge, 2007.

[40] J. H. Bramble and S. R. Hilbert, Estimation of linear functionals on Sobolev spaces with application to Fourier transforms and spline interpolation, *SIAM J. Numer. Anal.* **7** (1970), 113–124.

[41] O. Bratteli and P. Jorgensen, *Wavelets through a Looking Glass: The World of the Spectrum*, Birkhäuser, Boston, 2002.

[42] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, third edition, Springer-Verlag, New York, 2008.

[43] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, Berlin, 1991.

[44] A. M. Bruckner, J. B. Bruckner and B. S. Thomson, *Real Analysis*, Prentice-Hall, New Jersey, 1997.

[45] C. Canuto, M. Y. Hussaini, A. Quarteroni and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.

[46] C. Canuto and A. Quarteroni, Approximation results for orthogonal polynomials in Sobelov spaces, *Math. Comput.* **38** (1982), 67–86.

[47] J. Céa, Approximation variationnelle des problémes aux limites, *Ann. Inst. Fourier (Grenoble)* **14** (1964), 345–444.

[48] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.

[49] W. Cheney, *Analysis for Applied Mathematics*, Springer-Verlag, New York, 2001.

[50] M. Chipot, *Variational Inequalities and Flow in Porous Media*, Springer-Verlag, New York, 1984.

[51] C. Chui, *An Introduction to Wavelets*, Academic Press, New York, 1992.

[52] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North Holland, Amsterdam, 1978.

[53] P. G. Ciarlet, Basic error estimates for elliptic problems, in P. G. Ciarlet and J.-L. Lions, eds., *Handbook of Numerical Analysis*, Vol. II, North-Holland, Amsterdam, 1991, 17–351.

[54] P. Clément, Approximation by finite element functions using local regularization, *RAIRO Anal. Numer.* **9R2** (1975), 77–84.

[55] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., *Discontinuous Galerkin Methods. Theory, Computation and Applications*, Lecture Notes in Comput. Sci. Engrg. **11**, Springer-Verlag, New York, 2000.

[56] L. Collatz, *Functional Analysis and Numerical Mathematics*, Academic Press, New York, 1966.

[57] D. Colton, *Partial Differential Equations: An Introduction*, Random House, New York, 1988.

[58] J. Conway, *A Course in Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1990.

[59] J. W. Cooley and J. W. Tukey, An algorithm for the machine calculations of complex Fourier series, *Math. Comp.* **19** (1965), 297–301.

[60] C. Cryer, *Numerical Functional Analysis*, Clarendon Press, 1982.

[61] B. Dacorogna, *Direct Methods in the Calculus of Variations*, second edition, Springer, 2008.

[62] I. Daubechies, Orthonormal bases of compactly supported wavelets, *Comm. Pure Appl. Math.* **41** (1988), 909–996.

[63] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[64] P. Davis, *Interpolation and Approximation*, Blaisdell, New York, 1963.

[65] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, 1978.

[66] F. de Hoog and R. Weiss, Asymptotic expansions for product integration, *Math. of Comp.* **27** (1973), 295–306.

[67] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[68] P. Deuflhard, *Newton Methods for Nonlinear Problems, Affine Invariance and Adaptive Algorithms*, Springer, Berline, 2004.

[69] F. Deutsch, *Best Approximation in Inner Product Spaces*, CMS Books in Mathematics, Vol. 7, Springer-Verlag, New York, 2001.

[70] S. Drabla, M. Sofonea and B. Teniou, Analysis of some frictionless contact problems for elastic bodies, *Ann. Pol. Mat.* **LXIX** (1998), 75–88.

[71] N. Dunford and J. Schwartz, *Linear Operators. Part I: General Theory*, Interscience Pub., New York, 1964.

[72] C. Dunkl and Y. Xu, *Orthogonal Polynomials of Several Variables*, Cambridge Univ. Press, Cambridge, 2001.

[73] J. Duoandikoetxea, *Fourier Analysis*, Graduate Studies in Mathematics, Vol. 29, American Mathematical Society, Providence, 2001.

[74] G. Duvaut and J.-L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[75] R. Edwards, *Functional Analysis*, Holt, Rinehart and Winston, New York, 1965.

[76] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[77] P. Enflo, A counterexample to the approximation problem in Banach spaces, *Acta Mathematica* **130** (1973), 309–317.

[78] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, 1998.

[79] R. S. Falk, Error estimates for the approximation of a class of variational inequalities, *Math. Comp.* **28** (1974), 963–971.

[80] I. Fenyö and H. Stolle, *Theorie und Praxis der linearen Integralgleichungen - 2*, Birkhäuser-Verlag, 1983.

[81] G. Fichera, Problemi elastostatici con vincoli unilaterali: il problema di Signorini con ambigue condizioni al contorno, *Mem. Accad. Naz. Lincei* **8** (7) (1964), 91–140.

[82] J. Flores, The conjugate gradient method for solving Fredholm integral equations of the second kind, *Intern. J. Computer Math.* **48** (1993), 77–94.

[83] E. Foufoula-Georgiou and P. Kumar, eds., *Wavelets in Geophysics*, Academic Press, Inc., San Diego, 1994.

[84] J. Franklin, *Methods of Mathematical Economics*, Springer-Verlag, New York, 1980.

[85] A. Friedman, *Variational Principles and Free-boundary Problems*, John Wiley, New York, 1982.

[86] W. Freeden, T. Gervens, and M. Schreiner, *Constructive Approximation on the Sphere, with Applications to Geomathematics*, Oxford Univ. Press, 1998.

[87] R. Freund, G. H. Golub, and N. Nachtigal, Iterative solution of linear systems, *Acta Numerica—1992*, Cambridge University Press, pp. 57–100.

[88] W. Gautschi, Orthogonal polynomials: applications and computations, in *Acta Numerica—1996*, A. Iserles, ed., Cambridge University Press, Cambridge, 1996, pp. 45–119.

[89] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 2001.

[90] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations, Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[91] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, 1984.

[92] R. Glowinski, *Finite Element Methods for Incompressible Viscous Flow*, *Handbook of Numerical Analysis*, Vol. IX, eds. P.G. Ciarlet and J.L. Lions, North-Holland, Elsevier, Amsterdam, 2003.

[93] R. Glowinski, J.-L. Lions and R. Trémolières, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[94] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer-Verlag, New York, 1996.

[95] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd edition, The Johns Hopkins University Press, Baltimore, 1996.

[96] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.

[97] I. Graham, Singularity expansions for the solution of second kind Fredholm integral equations with weakly singular convolution kernels, *J. Integral Eqns.* **4** (1982), 1–30.

[98] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[99] C. Groetsch, *Inverse Problems in the Mathematical Sciences*, Vieweg Pub., Braunschweig/Wiesbaden, 1993.

[100] T. Gronwall, On the degree of convergence of Laplace's series, *Transactions of the Amer. Math. Soc.* **15** (1914), 1–30.

[101] B. Gustafsson, *High Order Difference Methods for Time Dependent PDE*, Springer-Verlag, Berlin, Heidelberg, 2008.

[102] D. M. Ha, *Functional Analysis: I. A Gentle Introduction*, Matrix Editions, Ithaca, NY, 2006.

[103] W. Hackbusch, Die schnelle Auflösung der Fredholmschen Integralgleichungen zweiter Art, *Beiträge Numerische Math.* **9** (1981), 47–62.

[104] W. Hackbusch, *Multi-grid Methods and Applications*, Springer-Verlag, 1985.

[105] W. Hackbusch, *Integral Equations: Theory and Numerical Treatment*, Birkhäuser Verlag, Basel, 1994.

[106] W. Hackbusch and Z. Nowak, On the fast matrix multiplication in the boundary element method by panel clustering, *Numer. Math.* **54** (1989), 463–491.

[107] C. A. Hall and T. A. Porsching, *Numerical Analysis of Partial Differential Equations*, Prentice Hall, New Jersey, 1990.

[108] W. Han, The best constant in a trace inequality, *Journal of Mathematical Analysis and Applications* **163** (1992), 512–520.

[109] W. Han, Finite element analysis of a holonomic elastic-plastic problem, *Numer. Math.* **60** (1992), 493–508.

[110] W. Han, *A Posteriori Error Analysis via Duality Theory: With Applications in Modeling and Numerical Approximations*, Springer, New York, 2005.

[111] W. Han, S. Jensen, and B. D. Reddy, Numerical approximations of internal variable problems in plasticity: error analysis and solution algorithms, *Numerical Linear Algebra with Applications* **4** (1997), 191–204.

[112] W. Han and B. D. Reddy, On the finite element method for mixed variational inequalities arising in elastoplasticity, *SIAM J. Numer. Anal.* **32** (1995), 1778–1807.

[113] W. Han and B. D. Reddy, *Plasticity: Mathematical Theory and Numerical Analysis*, Springer-Verlag, New York, 1999.

[114] W. Han, B. D. Reddy, and G. C. Schroeder, Qualitative and numerical analysis of quasistatic problems in elastoplasticity, *SIAM J. Numer. Anal.* **34** (1997), 143–177.

[115] W. Han and M. Sofonea, *Quasistatic Contact Problems in Viscoelasticity and Viscoplasticity*, American Mathematical Society and International Press, 2002.

[116] W. Han and L.-H. Wang, Non-conforming finite element analysis for a plate contact problem, *SIAM Journal on Numerical Analysis* **40** (2002), 1683–1697.

[117] O. Hansen, K. Atkinson, and D. Chien, On the norm of the hyperinterpolation operator on the unit disk and its use for the solution of the nonlinear Poisson equation, *IMA J. Numer. Anal.* **28** (2008), doi: 10.1093/imanum/drm052.

[118] G. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, 2^{nd} ed., Cambridge University Press, 1952.

[119] J. Haslinger, I. Hlaváček and J. Nečas, Numerical methods for unilateral problems in solid mechanics, in P. G. Ciarlet and J.-L. Lions, eds., *Handbook of Numerical Analysis*, Vol. IV, North-Holland, Amsterdam, 1996, 313–485.

[120] P. Henrici, *Applied and Computational Complex Analysis*, Vol. 3, John Wiley, New York, 1986.

[121] M. R. Hestenes, *Conjugate Direction Methods in Optimization*, Springer-Verlag, Berlin, 1980.

[122] M. R. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Stand.* **49** (1952), 409–436.

[123] E. Hewitt and R. E. Hewitt, The Gibbs–Wilbraham phenomenon: an episode in Fourier analysis, *Archive for History of Exact Sciences* **21** (1979), 129–160.

[124] N. Higham, *Accuracy and Stability of Numerical Algorithms*, second edition, SIAM, Philadelphia, 2002.

[125] E. Hille and J. Tamarkin, On the characteristic values of linear integral equations, *Acta Math.* **57** (1931), 1–76.

[126] I. Hlaváček, J. Haslinger, J. Nečas and J. Lovíšek, *Solution of Variational Inequalities in Mechanics*, Springer-Verlag, New York, 1988.

[127] H. Huang, W. Han, and J. Zhou, The regularization method for an obstacle problem, *Numer. Math.* **69** (1994), 155–166.

[128] V. Hutson and J. S. Pym, *Applications of Functional Analysis and Operator Theory*, Academic Press, London, 1980.

[129] E. Isaacson and H. Keller, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[130] M. Jaswon and G. Symm, *Integral Equation Methods in Potential Theory and Elastostatics*, Academic Press, 1977.

[131] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, 1987.

[132] G. Kaiser, *A Friendly Guide to Wavelets*, Birkhäuser, Boston, 1994.

[133] S. Kakutani, Topological properties of the unit sphere of Hilbert space, *Proc. Imp. Acad. Tokyo* **19** (1943), 269–271.

[134] L. Kantorovich, Functional analysis and applied mathematics, *Uspehi Mat. Nauk* **3** (1948), 89–185.

[135] L. Kantorovich and G. Akilov, *Functional Analysis in Normed Spaces*, 2^{nd} ed., Pergamon Press, New York, 1982.

[136] L. Kantorovich and V. Krylov, *Approximate Methods of Higher Analysis*, Noordhoff, Groningen, 1964.

[137] H. Kardestuncer and D. H. Norrie (editors), *Finite Element Handbook*, McGraw-Hill Book Company, New York, 1987.

[138] F. Keinert, *Wavelets and Multiwavelets*, Chapman & Hall/CRC, Boca Raton, 2004.

[139] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[140] C. T. Kelley, *Iterative Methods for Optimization*, SIAM, Philadelphia, 1999.

[141] O. Kellogg, *Foundations of Potential Theory*, reprinted by Dover Pub, 1929.

[142] S. Kesavan, *Topics in Functional Analysis and Applications*, John Wiley, New Delhi, 1989.

[143] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM, Philadelphia, 1988.

[144] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and their Applications*, Academic Press, New York, 1980.

[145] T. Koornwinder, Two–variable analogues of the classical orthogonal polynomials, *Theory and Application of Special Functions*, ed. R. A. Askey, Academic Press, New York, 1975, 435–495.

[146] V. A. Kozlov, V. G. Maz'ya and J. Rossmann, *Elliptic Boundary Value Problems in Domains with Point Singularities*, American Mathematical Society, 1997.

[147] M. A. Krasnosel'skii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, New York, 1964.

[148] M. Krasnosel'skii and P. Zabreyko, *Geometric Methods of Nonlinear Analysis*, Springer-Verlag, 1984.

[149] R. Kress, *Linear Integral Equations*, second edition, Springer-Verlag, Berlin, 1998.

[150] R. Kress, *Numerical Analysis*, Springer, New York, 1998.

[151] R. Kress and I. Sloan, On the numerical solution of a logarithmic integral equation of the first kind for the Helmholtz equation, *Numer. Math.* **66** (1993), 199–214.

[152] S. Kumar and I. Sloan, A new collocation-type method for Hammerstein equations, *Math. Comp.* **48** (1987), 585–593.

[153] L. P. Lebedev and M. J. Cloud, *The Calculus of Variations and Functional Analysis, with Optimal Control and Applications in Mechanics*, World Scientific, Singapore, 2003.

[154] V. I. Lebedev, *An Introduction to Functional Analysis in Computational Mathematics*, Birkhäuser, Boston, 1997.

[155] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2^{nd} ed., Birkhäuser, Boston, 1992.

[156] D. Luenberger, *Linear and Nonlinear Programming*, 2^{nd} ed., Addison-Wesley, Reading, Mass., 1984.

[157] P. Linz, *Theoretical Numerical Analysis, An Introduction to Advanced Techniques*, John Wiley, New York, 1979.

[158] J.-L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications*, three volumes, Springer-Verlag, New York, 1968.

[159] J.-L. Lions and G. Stampacchia, Variational inequalities, *Comm. Pure Appl. Math.* **20** (1967), 493–519.

[160] B. Logan and L. Shepp, Optimal reconstruction of a function from its projections, *Duke Math. Journal* **42** (1975), pp. 645-659.

[161] T. MacRobert, *Spherical Harmonics*, 3^{rd} ed., Pergamon Press, 1967.

[162] Y. Maday and A. Quarteroni, Legendre and Chebyshev spectral approximations of Burgers' equation, *Numer. Math.* **37** (1981), 321–332.

[163] S. Mallat, A theory for multi-resolution approximation: the wavelet approximation, *IEEE Trans. PAMI* **11** (1989), 674–693.

[164] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.

[165] G. I. Marchuk, Splitting and alternating direction methods, in P. G. Ciarlet and J.-L. Lions, eds., *Handbook of Numerical Analysis*, Vol. I, North-Holland, Amsterdam, 1990, 197–464.

[166] G. I. Marchuk and V. V. Shaidurov, *Difference Methods and Their Extrapolations*, Springer-Verlag, New York, 1983.

[167] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, 2000.

[168] R. McOwen, *Partial Differential Equations, Methods and Applications*, second edition, Prentice Hall, NJ, 2003.

[169] G. Meinardus, *Approximation of Functions: Theory and Numerical Methods*, Springer-Verlag, New York, 1967.

[170] Y. Meyer, *Wavelets and Operators*, Cambridge Studies in Advanced Mathematics, **37**, Cambridge University Press, 1992.

[171] S. Mikhlin, *Integral Equations*, 2^{nd} ed., Pergamon Press, 1964.

[172] S. Mikhlin, *Mathematical Physics: An Advanced Course*, North-Holland Pub., 1970.

[173] L. Milne-Thomson, *Theoretical Hydrodynamics*, 5^{th} ed., Macmillan, New York, 1968.

[174] G. J. Minty, Monotone (non linear) operators in Hilbert spaces, *Duke Math. J.* **29** (1962), 341–346.

[175] R. E. Moore, *Computational Functional Analysis*, Ellis Horwood Limited, Chichester, 1985.

[176] J. Nečas, Sur une méthode pour resoudre les equations aux derivées partielles du type elliptique, voisine de la variationnelle, *Ann. Scuola Norm. Sup. Pisa* **16** (1962), 305–326.

[177] J. Nečas and I. Hlavaček, *Mathematical Theory of Elastic and Elastoplastic Bodies: An Introduction*, Elsevier, Amsterdam, 1981.

[178] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser, Basel, 1993.

[179] R. A. Nicolaides, On a class of finite elements generated by Lagrange interpolation, *SIAM J. Numer. Anal.* **9** (1972), 435–445.

[180] Y. Nievergelt, *Wavelets Made Easy*, Birkhäuser, Boston, 2001.

[181] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 1999.

[182] E. Nyström, Über die praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben, *Acta Math.* **54** (1930), 185–204.

[183] J. T. Oden and J. N. Reddy, *An Introduction to the Mathematical Theory of Finite Elements*, John Wiley, New York, 1976.

[184] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[185] P. D. Panagiotopoulos, *Inequality Problems in Mechanics and Applications*, Birkhäuser, Boston, 1985.

[186] W. Patterson, *Iterative Methods for the Solution of a Linear Operator Equation in Hilbert Space—A Survey*, Springer Lecture Notes in Mathematics, No. 394, Springer-Verlag, 1974.

[187] W. Pogorzelski, *Integral Equations and Their Applications*, Pergamon Press, 1966.

[188] C. Pozrikidis, *Boundary Integral and Singularity Methods for Linearized Viscous Flow*, Cambridge Univ. Press, Cambridge, 1992.

[189] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, New York, 1994.

[190] D. Ragozin, Constructive polynomial approximation on spheres and projective spaces, *Trans. Amer. Math. Soc.* **162** (1971), 157–170.

[191] D. Ragozin, Uniform convergence of spherical harmonic expansions, *Math. Annalen* **195** (1972), 87–94.

[192] B. D. Reddy, *Introductory Functional Analysis with Applications to Boundary Value Problems and Finite Elements*, Springer, New York, 1998.

[193] B. D. Reddy and T. B. Griffin, Variational principles and convergence of finite element approximations of a holonomic elastic-plastic problem, *Numer. Math.* **52** (1988), 101–117.

[194] J. Rice, On the degree of convergence of nonlinear spline approximation, in *Approximations with Special Emphasis on Spline Functions*, ed. by I.J. Schoenberg, Academic Press, New York, 1969, 349–365.

[195] T. Rivlin, *Chebyshev Polynomials*, 2^{nd} ed., John Wiley, New York, 1990.

[196] J. E. Roberts and J.-M. Thomas, Mixed and hybrid methods, in P. G. Ciarlet and J.-L. Lions, eds., *Handbook of Numerical Analysis*, Vol. II, North-Holland, Amsterdam, 1991, 523–639.

[197] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.

[198] H. Royden, *Real Analysis*, 3^{rd} ed., Collier-MacMillan, 1988.

[199] W. Rudin, *Real and Complex Analysis*, 3^{rd} ed., McGraw-Hill, New York, 1987.

[200] W. Rudin, *Functional Analysis*, 2^{nd} ed., McGraw-Hill, New York, 1991.

[201] K. Rustamov, On approximation of functions on the sphere, *Russian Acad. Sci. Izv. Math.* **43** (1994), 311–329.

[202] C. Schneider, Regularity of the solution to a class of weakly singular Fredholm integral equations of the second kind, *Integral Eqns & Operator Thy* **2** (1979), 62–68.

[203] C. Schneider, Product integration for weakly singular integral equations, *Math. of Comp.* **36** (1981), 207–213.

[204] C. Schwab, *p- and hp-Finite Element Methods*, Oxford University Press, 1998.

[205] I. Sloan, Improvement by iteration for compact operator equations, *Math. Comp.* **30** (1976), 758–764.

[206] I. Sloan, Error analysis of boundary integral methods, *Acta Numerica* **1** (1992), 287–339.

[207] I. Sloan and R. Womersley, Constructive polynomial approximation on the sphere, *J. Approx. Theor.* **103** (2000), 91–118.

[208] M. Sofonea and A. Matei, *Variational Inequalities with Applications: A Study of Antiplane Frictional Contact Problems*, Springer, 2009.

[209] P. Šolín, K. Segeth, and I. Doležel, *Higher–Order Finite Element Methods*, Chapman & Hall/CRC, Boca Raton, 2004.

[210] I. Stakgold, *Green's Functions and Boundary Value Problems*, John Wiley, New York, 1979.

[211] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[212] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[213] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer-Verlag, New York, 1993.

[214] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[215] E. J. Stollnitz, T. D. Derose, and D. H. Salesin, *Wavelets for Computer Graphics: Theory and Applications*, Morgan Kaufmann Publishers, Inc., San Francisco, 1996.

[216] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[217] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley–Cambridge Press, MA, 1996.

[218] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks/Cole Advanced Books & Software, 1989.

[219] B. Szabó and I. Babuška, *Finite Element Analysis*, John Wiley, Inc., New York, 1991.

[220] G. Szegö, *Orthogonal Polynomials*, revised edition, American Mathematical Society, New York, 1959.

[221] V. Thomée, Finite difference methods for linear parabolic equations, in P. G. Ciarlet and J.-L. Lions, eds., *Handbook of Numerical Analysis*, Vol. I, North-Holland, Amsterdam, 1990, 5–196.

[222] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer, 1997.

[223] L. Trefethen, *Spectral Methods in* MATLAB, SIAM Pub., Philadelphia, 2000.

[224] L. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM Pub., Philadelphia, 1997.

[225] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland Publ. Comp., Amsterdam, 1978.

[226] C. Van Loan, *Computational Frameworks for the Fast Fourier Transform*, SIAM, 1992.

[227] C. R. Vogel, *Computational Methods for Inverse Problems*, SIAM, Philadelphia, 2002.

[228] L.-H. Wang, Nonconforming finite element approximations to the unilateral problem, *Journal of Computational Mathematics* **17** (1999), 15–24.

[229] W. Wendland, Boundary element methods and their asymptotic convergence, in *Theoretical Acoustics and Numerical Techniques*, ed. by P. Filippi, Springer-Verlag. CISM Courses and Lectures No. 277, International Center for Mechanical Sciences, 1983.

[230] W. Wendland, Strongly elliptic boundary integral equations, in *The State of the Art in Numerical Analysis*, ed. by A. Iserles and M. Powell, Clarendon Press, 1987, 511–562.

[231] W. Wendland, Boundary element methods for elliptic problems, in *Mathematical Theory of Finite and Boundary Element Methods*, ed. by A. Schatz, V. Thomée, and W. Wendland, Birkhäuser, Boston, 1990, pp. 219–276.

[232] R. Winther, Some superlinear convergence results for the conjugate gradient method, *SIAM J. Numer. Anal.* **17** (1980), 14–17.

[233] J. Wloka, *Partial Differential Equations*, Cambridge University Press, UK, 1987.

[234] P. Wojtaszczyk, *A Mathematical Introduction to Wavelets*, London Mathematical Society Student Texts **37**, Cambridge University Press, 1997.

[235] J. Xu, Iterative methods by space decomposition and subspace correction, *SIAM Review* **34** (1992), 581–613.

[236] J. Xu and L. Zikatanov, Some observations on Babuška and Brezzi theories, *Numerische Mathematik* **94** (2003), 195–202.

[237] Y. Xu, Orthogonal polynomials and cubature formulae on spheres and balls, *SIAM J. Math. Anal.* **29** (1998), 779–793.

[238] Y. Xu, Representation of reproducing kernels and the Lebesgue constants on the ball, *J. Approx. Theor.* **112** (2001), 295–310.

[239] Y. Xu, Lecture notes on orthogonal polynomials of several variables, in *Advances in the Theory of Special Functions and Orthogonal Polynomials*, ed. by W. zu Castell, F. Filbir, and B. Forster, Nova Science Publishers, 2005, pp. 141–196.

[240] Y. Xu, Weighted approximation of functions on the unit sphere, *Const. Approx.* **21** (2005), 1–28.

[241] Y. Xu, Analysis on the unit ball and on the simplex, *Elec. Trans. Numer. Anal.* **25** (2006), 284–301.

[242] Y. Yan and I. Sloan, On integral equations of the first kind with logarithmic kernels, *J. Integral Eqns & Applics* **1** (1988), 517–548.

[243] N. N. Yanenko, *The Method of Fractional Steps*, Springer-Verlag, New York, 1971.

[244] E. Zeidler, *Nonlinear Functional Analysis and its Applications. I: Fixed-point Theorems*, Springer-Verlag, New York, 1985.

[245] E. Zeidler, *Nonlinear Functional Analysis and its Applications. II/A: Linear Monotone Operators*, Springer-Verlag, New York, 1990.

[246] E. Zeidler, *Nonlinear Functional Analysis and its Applications. II/B: Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

[247] E. Zeidler, *Nonlinear Functional Analysis and its Applications. III: Variational Methods and Optimization*, Springer-Verlag, New York, 1986.

[248] E. Zeidler, *Nonlinear Functional Analysis and its Applications. IV: Applications to Mathematical Physics*, Springer-Verlag, New York, 1988.

[249] E. Zeidler, *Applied Functional Analysis: Applications of Mathematical Physics*, Springer-Verlag, New York, 1995.

[250] E. Zeidler, *Applied Functional Analysis: Main Principles and Their Applications*, Springer-Verlag, New York, 1995.

[251] A. Zygmund, *Trigonometric Series*, Vols. I and II, Cambridge Univ. Press, 1959.

Index

$B(\cdot, r)$, 10
$C(\Omega)$, 2
$C(\overline{\Omega})$, 3
C^k domain, 283
$C^m(\Omega)$, 3, 40
$C^m(\overline{\Omega})$, 3, 40
$C^\infty(\Omega)$, 41
$C^\infty(\overline{\Omega})$, 41
$C^{m,\beta}(\overline{\Omega})$, 42
$C_p^k(2\pi)$, 3
$C_0^\infty(\Omega)$, 41
$C_p(2\pi)$, 3
$H^k(\Omega)$, 284
    inner product, 286
$H^s(2\pi)$, 312
$H^s(\Omega)$, 290
    inner product, 290
$H^s(\partial\Omega)$, 292
$H_0^s(\Omega)$, 291
$H^{-1}(\Omega)$, 291
$H^m(a, b)$, 28
$H_0^k(\Omega)$, 290
$L^p(\Omega)$, 44
$L_w^p(\Omega)$, 18
$L_w^\infty(\Omega)$, 18
$W^{k,p}(\Omega)$, 284
    norm, 284
    seminorm, 285
$W^{m,p}(a, b)$, 19
$W^{s,p}(\Omega)$, 290
    norm, 290
$W^{s,p}(\partial\Omega)$, 292
$W_0^{s,p}(\Omega)$, 291
$W_0^{k,p}(\Omega)$, 290
Π_n^d, 584
$\mathbb{P}_k$, 6
$\mathbb{P}_n^d$, 585
$\mathbb{R}^d$, 349
$\mathbb{R}_+$, 438
$\mathbb{S}^d$
    inner product, 349
    norm, 349
$\mathbb{V}_n^d$, 585
$\mathbb{Z}$, 40
$\mathbb{Z}_+$, 40
$\mathcal{L}(V, W)$, 59
$\overline{\mathbb{R}}$, 430
∂^α, 39
$r_\sigma(B)$, 58
meas(D), 16

a posteriori error estimate, 451
a.e., 17
Abel integral operator, 108
abstract interpolation problem, 118
Arzela-Ascoli Theorem, 50
associated Legendre function, 320
Aubin-Nitsche Lemma, 417

Babuška-Brezzi condition, 376
backward triangle inequality, 10
Banach fixed-point theorem, 208, 209
Banach space, 15
  uniformly convex, 94
Banach-Steinhaus theorem, 75
barycentric coordinates, 396
basis, 5, 402
  countably-infinite, 13
  orthogonal, 29
  orthonormal, 29
  Schauder, 13
Bessel's inequality, 29
best approximation, 131
  existence, 137
  inner product space, 142
  trigonometric polynomial, 139, 146
  uniqueness, 138
best constant, 306, 307
best uniform approximation, 138
bidual, 92
biharmonic equation, 348
bijection, 53
bilinear form
  V-elliptic, 334
  bounded, 334
boundary integral equation, 551
  direct type, 559
  first kind, 577
  indirect type, 562
  numerical, first kind, 579
  numerical, second kind, 565
  second kind, 565
boundary value problem
  Adler, 345
  homogeneous Dirichlet, 338
  linearized elasticity, 348
  Neumann, 341
  non-homogeneous Dirichlet, 339
  Robin, 345
Brouwer's fixed-point theorem, 241
bulk modulus, 351

Cartesian product, 6
Cauchy sequence, 14
Céa's inequality, 372
characteristic function, 16
Chebyshev Equi-oscillation Theorem, 139
Chebyshev polynomial, 151
Christoffel-Darboux Identity, 163
circulant matrix, 148
Clarkson inequality, 46
closed operator, 330
closed range theorem, 332
closed set, 10, 132
  weakly, 132
collocation
  iterated, 498
  piecewise linear, 483
  trigonometric polynomial, 486
compact linear operator, 95
compact operator, 95
  $C(D)$, 96
  $L^2(a,b)$, 99
  properties, 97
compact set, 49
compact support, 41
compactness
  weak, 92
completely continuous vector fields, 241, 243
  rotation, 243
completion, 15
condensation technique, 390

condition number, 74
cone, 438
conjugate gradient iteration
    convergence, 246
    superlinear convergence, 247
conjugate gradient method, 245
    operator equation, 245
    variational formulation, 378
consistency, 68, 263
constitutive law, 349
contact problem in elasticity
    frictional, 460
    frictionless with deformable support, 469
    Signorini frictionless, 465
continuity, 10
contractive mapping theorem, 210
convergence, 10, 266
    strong, 90
    weak, 91
    weak-*, 93
convergence order, 267
convex combination, 132
convex function, 132
    strictly, 132
convex minimization, 231
    Gâteaux derivative, 231
convex set, 132
countably-infinite basis, 13

Daubechies wavelets, 203
    scaling function, 202
degenerate kernel function, 98
dense set, 13
density argument, 324
derivatives
    mean value theorems, 229
    properties, 227
difference
    backward, 254
    centered, 254
    forward, 254
differential calculus, 225
differential equation
    Banach space, 221
dimension, 5
Dirac δ-function, 85
Dirichlet kernel function, 160
discrete Fourier transform, 187
discrete Galerkin method, 593
discrete orthogonal projection operator, 593
displacement, 348
divergence theorem, 553
domain
    operator, 52
    spatial, 39
double layer potential, 562
    evaluation, 568
double obstacle problem, 440
dual problem, 357
dual space, 79
dyadic interval, 193
dyadic number, 193

eigenvalue, 105
eigenvector, 105
elastic plate problems, 348
elasticity tensor, 350
elasto-plastic torsion problem, 439
elliptic variational inequality
    existence, 428, 431
    first kind, 434
    second kind, 434
    uniqueness, 428, 431
epigraph, 433
equivalence class, 17
equivalent norms, 11
exterior Dirichlet problem, 556, 562
exterior Neumann problem, 557, 560

fast Fourier transform (FFT), 188
finite difference method, 253
    convergence, 266
    explicit, 255

implicit, 255
stability, 266
finite element method, 383
h-p-version, 421
h-version, 421
p-version, 421
convergence, 415
error estimate, 415
finite element space, 404
finite elements
affine-equivalent, 402
regular family, 411
first kind integral equation, 577
fixed point, 207
Fourier coefficients, 168
Fourier series, 146, 168
coefficients, 168
complex form, 169
partial sum, 160
uniform error bound, 161
Fourier transform
discrete, 187
Fréchet derivative, 225
Fréchet differentiable, 226
Fredholm alternative theorem, 101
frequency, 168
Fubini's theorem, 17
functionals
linearly independent, 118

Galerkin method, 368
convergence, 372
generalized, 377
iterated, 497
piecewise linear, 488
trigonometric polynomial, 490
uniform convergence, 491
Gâteaux derivative, 225
Gâteaux differentiable, 226
Gauss-Seidel method, 216
generalized Galerkin method, 377
generalized Lax-Milgram Lemma, 359
generalized solution, 261
generalized variational lemma, 278
geometric series theorem, 61
generalization, 64
Gibbs phenomenon, 173
Gram-Schmidt method, 34
Green's first identity, 554
Green's representation formula, 554
Green's representation formula on exterior region, 558
Green's second identity, 554
Gronwall's inequality, 224

Haar scaling spaces, 193
Haar wavelet function, 195
Haar wavelet spaces, 195
Haar wavelets, 192
scaling function, 192
Hahn-Banach Theorem, 82
Hammerstein integral equation, 218
heat equation, 254
Heine-Borel Theorem, 49
Helmholtz equation, 338
Hencky material, 459
Hermite finite element, 402
Hermite polynomial interpolation, 122
Hilbert space, 27
Hilbert-Schmidt integral operator, 250
Hilbert-Schmidt kernel function, 100
Hölder continuous, 42
Hölder inequality, 45
Hölder space, 41
homotopy, 244
hyperinterpolation operator, 593

ill-conditioned, 75
ill-posed, 75

index of a fixed point, 244
inf-sup condition, 376
injective, 53
inner product, 22
integral equation, 216
  collocation method, 474
  Galerkin method, 476
  Hammerstein, 218
  iteration, 218
  nonlinear, 217
  Nyström method, 504
  projection method, 474
  Urysohn, 217, 547
  Volterra, 218
integral operator
  Abel, 108
  self-adjoint, 88
integration by parts formula, 323
interior Dirichlet problem, 552, 559
interior Neumann problem, 552, 559
internal approximation method, 376
interpolation error estimate
  global, 412
  local, 409
  over the reference element, 408
interpolation operator, 407
interpolation projection, 155
interpolation property, 46
interpolation theory, 118
interpolatory projection, 164
invariant subspace, 106
isomorphic, 6
iteration method
  integral equations, 531
  Laplace's equation, 215, 539
  Nyström method, 532

Jackson's Theorem, 159
  multivariable polynomial approximation, 585
Jacobi method, 216
Jacobi polynomial, 149

Kelvin transformation, 555
Korn's inequality, 301
Krylov subspace, 251

l.s.c., 430
Lagrange basis functions, 120
Lagrange finite element, 402
Lagrange polynomial interpolation, 120
Lagrange's formula, 120
Lagrangian multiplier, 451
Lamé moduli, 350
Laplace expansion, 320
Lax equivalence theorem, 266
Lax-Milgram Lemma, 335, 336
  generalized, 359
Lebesgue constant, 161
Lebesgue Dominated Convergence Theorem, 17
Lebesgue integral, 17
Lebesgue integration, 16
Lebesgue measure, 16
Legendre polynomial, 34, 150, 319
  normalized, 145
linear algebraic system, 214
linear combination, 4
linear function, 6
linear functional, 79
  extension, 80
linear integral operator, 62
linear interpolant, 408
linear space, 2
linear system
  iterative method, 214
linearized elasticity, 348
linearly dependent, 4
linearly independent, 4
Lipschitz constant, 42
Lipschitz continuous, 41
Lipschitz domain, 283
load vector, 368

local truncation error, 270
locally p-integrable, 278
lower semicontinuous (l.s.c.) function, 133

mapping, 52
material
  anisotropic, 350
  homogeneous, 350
  isotropic, 350
  nonhomogeneous, 350
Mazur Lemma, 136
measurable function, 17
mesh parameter, 411
mesh refinement, 419
meshsize, 394
method of Lagrangian multiplier, 451
minimal angle condition, 414
Minkowski inequality, 45
Minty Lemma, 435
mixed formulation, 358
modulus of continuity, 124
monomial, 39
multi-index notation, 39
multiresolution analysis, 199

Nekrasov's equation, 218
Newton's method, 236
  convergence, 236
  Kantorovich theorem, 238
  nonlinear differential equation, 240
  nonlinear integral equation, 239
  nonlinear system, 239
node, 396, 401
nonlinear algebraic equation, 213
nonlinear equation, 361
  Newton's method, 236
  projection method, 542
nonlinear functional analysis, 207
nonlinear integral equation, 239
nonlinear operator
  completely continuous, 242
  derivatives, 225
  Taylor approximation, 243
norm, 7
  operator, 57
normed space, 8
  reflexive, 92
numerical quadrature
  convergence, 76
Nyström method
  asymptotic error formula, 514
  collectively compact operator approximation theory, 516
  conditioning, 515
  error analysis, 507
  iteration method, 532
  product integration, 518

obstacle problem, 424
open mapping theorem, 74
open set, 10
operator, 52
  addition, 52
  adjoint, 85
  bounded, 54
  closed, 330
  compact, 95
  compact linear, 95
  completely continuous, 95
  continuous, 54
  continuous linear, 55
  contractive, 209
  differentiation, 53
  extension, 72, 80
  finite rank, 97
  linear, 55
  linear integral, 62
  Lipschitz continuous, 209, 430
  non-expansive, 209
  norm, 57
  perturbation, 66
  projection, 143

scalar multiplication, 52
self-adjoint, 87
strongly monotone, 430
orthogonal, 28
orthogonal basis, 29
orthogonal projection, 155, 156
orthonormal basis, 29
orthonormal system, 29

Parallelogram Law, 25
parameter, 401
Parseval's equality, 31
Petrov-Galerkin method, 374
convergence, 376
Picard iteration, 210
Picard-Lindelöf Theorem, 221
piecewise polynomial interpolation, 124
piecewise polynomial method, 483
Plancherel formula, 186
Poincaré-Friedrichs inequality, 301
Poincaré-Wirtinger inequality, 305
point evaluation, 81
Poisson equation, 328
Poisson's ratio, 351
polarization identity, 25
polynomial approximation
relation to trigonometric polynomial approximation, 158
polynomial interpolation
barycentric formula, 128
error formula, 121
polynomial invariance property, 409
primal formulation, 354
primal problem, 357
principle of uniform boundedness, 75
product integration
error analysis, 520
graded mesh, 525
projection
interpolation, 155
on closed convex set, 143
orthogonal, 155
projection method, 474
conditioning, 492
error bound, 480
finite problem, 547
homotopy argument, 545
iterated, 494
theory, 477
projection operator, 154
on closed convex set, 143
proper functional, 430

quasiuniform triangulation, 414

range, 52
reference element, 398
reference element technique, 391, 410
reflexive normed space, 92
regularity condition, 394
regularization technique, 449
reproducing kernel function, 163
resolvent operator, 109
resolvent set, 109
Riesz representation theorem, 82
Ritz method, 369
Ritz-Galerkin method, 369
rotation, 545
rotation of a completely continuous vector field, 243

saddle point problem, 355
scaling equation, 201
Schauder basis, 13
Schauder's fixed-point theorem, 243
scheme
backward, 255
backward-time centered-space, 255
Crank-Nicolson, 256

forward, 255
forward-time centered-space, 255
generalized mid-point, 275
Lax-Wendroff, 259
two-level, 269
Schwartz space, 182
Schwarz inequality, 23
semi-norm, 8
separable space, 13
Signorini problem, 440
Simpson's rule, 323
sine integral function, 174
single layer potential, 562
evaluation, 572
singular elements, 419
smooth test function, 328
Sobolev quotient space, 302
Sobolev space, 19
compact embedding, 296
density result, 294
dual, 314
embedding, 295, 315
extension, 294
Fourier transform characterization, 308
norm equivalence theorem, 298
of integer order, 284
of real order, 290
over boundary, 292
periodic, 311
trace, 297
solution operator, 260
SOR method, 216
space of continuously differentiable functions, 39
span, 5
spectral method, 483
spectral radius, 215
spectrum, 109
continuous, 110
point, 110
residual, 110
spherical harmonics, 107, 318
spherical polynomial, 318
stability, 68, 266
stability estimate, 331
stable, 75
stiffness matrix, 368
condition number, 415
Stokes equations, 358
Stone–Weierstrass Theorem, 116
strain tensor, 349
stress tensor, 348
strictly convex function, 132
strictly normed space, 140
strong convergence, 90
strongly monotone, 331
strongly monotone operator, 430
sublinear, 82
sublinear functional, 82
subspace, 3
superconvergence, 502
surjective, 53

Toeplitz matrix, 148
transformation, 52
trapezoidal rule, 316
Tresca's law, 468
triangle inequality, 8
trigonometric interpolation, 126
convergence, 164
triple recursion relation, 152
Chebyshev polynomials, 151
Legendre polynomials, 150
multivariable, 588

uniform error bound, 157
uniform integrability, 93
unstable, 75
Urysohn integral equation, 217

vector space, 2
Volterra integral equation, 218
second kind, 64

wavelets, 191
Daubechies, 203

Haar, 192
weak compactness, 92
weak convergence, 91
weak derivative, 278
weak formulation, 329
weakly closed set, 132
weakly lower semicontinuous (w.l.s.c.) function, 133
weakly sequentially lower semicontinuous function, 133
Weierstrass Theorem, 116
well-conditioned, 75
well-posed, 75, 260

Young's inequality, 44
modified, 45
Young's modulus, 351